

Jian Yao  
Yang Xiao  
Peng You  
Guang Sun *Editors*

# The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)

# Lecture Notes in Electrical Engineering

## Volume 813

### Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy

Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico

Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India  
Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany

Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China

Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany

Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China

Gianluigi Ferrari, Università di Parma, Parma, Italy

Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain

Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany

Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China  
Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt

Torsten Kroeger, Stanford University, Stanford, CA, USA

Yong Li, Hunan University, Changsha, Hunan, China

Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA  
Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore

Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany

Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA

Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany

Subhas Mukhopadhyay, School of Engineering & Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand

Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA

Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi “Roma Tre”, Rome, Italy

Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Gan Woon Seng, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany

Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal

Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China

Walter Zamboni, DIEM - Università degli studi di Salerno, Fisciano, Salerno, Italy

Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact [leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com).

To submit a proposal or request further information, please contact the Publishing Editor in your country:

#### **China**

Jasmine Dou, Editor ([jasmine.dou@springer.com](mailto:jasmine.dou@springer.com))

#### **India, Japan, Rest of Asia**

Swati Meherishi, Editorial Director ([Swati.Meherishi@springer.com](mailto:Swati.Meherishi@springer.com))

#### **Southeast Asia, Australia, New Zealand**

Ramesh Nath Premnath, Editor ([ramesh.premnath@springernature.com](mailto:ramesh.premnath@springernature.com))

#### **USA, Canada:**

Michael Luby, Senior Editor ([michael.luby@springer.com](mailto:michael.luby@springer.com))

#### **All other Countries:**

Leontina Di Cecco, Senior Editor ([leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com))

**\*\* This series is indexed by EI Compendex and Scopus databases. \*\***

More information about this series at <https://link.springer.com/bookseries/7818>

Jian Yao · Yang Xiao · Peng You · Guang Sun  
Editors

# The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)



*Editors*

Jian Yao

School of Remote Sensing and  
Information Engineering  
Wuhan University  
Wuhan, Hubei, China

Peng You

National University of  
Defense Technology  
Changsha, Hunan, China

Yang Xiao

Department of Computer Science  
University of Alabama  
Tuscaloosa, AL, USA

Guang Sun

Big Data Institute  
Hunan University of Finance  
and Economics  
Changsha, Hunan, China

ISSN 1876-1100

Lecture Notes in Electrical Engineering

ISBN 978-981-16-6962-0

<https://doi.org/10.1007/978-981-16-6963-7>

ISSN 1876-1119 (electronic)

ISBN 978-981-16-6963-7 (eBook)

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.  
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,  
Singapore

# Preface

With the rapid development of transportation, finance, science and education, information, medical treatment, economy and trade, tourism and other aspects, the integration of industrialization and informatization deepens, and system intelligence has become the hot spot and mainstream of development in today's information society. As its foundation and key technologies, computer vision, image processing, pattern recognition and so on are the research directions that the global industry is competing to invest in.

The International Conference on Image, Vision and Intelligent Systems (ICIVIS) was held in Changsha, China, from June 18–20, 2021, which aims to provide a professional and efficient communication platform for researchers and scholars from all over the world. We will jointly discuss the academic trends and development trends in various directions in the fields of images, vision and intelligence systems and related research fields, discuss the current hot issues, share research results, promote the development and progress of related research and applications, promote the development of disciplines and promote personnel training.

## ICIVIS 2021

Approximately 110 participants of conference came from various research institutions and universities such as Beijing University of Technology, China Agricultural University, Jinan University, China University of Petroleum, National University of Defense Technology, Central South University, Huazhong University of Science and Technology, Chongqing University of Posts and Telecommunications, Jiangsu Institute of Automation, Central China Institute of Optoelectronics, Information Engineering University, Air Force Engineering University, Shanghai University, Changsha University of Technology, Central South University of Forestry Science and Technology, Shanghai Institute of Aerospace Control Technology, Guizhou University, Anhui University of Technology, Nanchang College of Engineering,

Hubei University of Technology, Fujian University of Agriculture and Forestry, Beijing Institute of Materials, Henan University and Chang'an University.



Group photo

ICIVIS 2021 was sponsored by Hunan University, National University of Defense Technology, Hunan City University, Hunan University of Finance and Economics and Xiangnan University and was held in Changsha, October 23–24, 2020. It is an annual forum dedicated to the emerging and challenging topics in intelligent computation and automation technology. We also highly appreciate the sponsors who financially supported this event.

Besides the sponsors, the success of relied on the big support from the committee as follows:

### **General Chairs**

Prof. Jian Yao, Wuhan University, China

Prof. Yang Xiao, University of Alabama, USA

Prof. Guang Sun, Hunan University of Finance and Economics, China

Prof. Peng You, National University of Defense Technology, China

Prof. Jinping Li, University of Jinan, China

### **Conference Program Chairs**

Prof. Hongwei Ge, Dalian University of Technology, China

Prof. Shengling Geng, Qinghai Normal University, China

Prof. Aiping Qu, University of South China, China

Prof. Xiushan Nie, Shandong Jianzhu University, China

Prof. Wei Wei, Xi'an University of Technology, China

**Technical Chairs**

- Prof. Lixin Tan, Hunan Agricultural University, China  
Assoc. Prof. Yunze He, Hunan University, China  
Prof. Huaiqing He, Civil Aviation University of China, China  
Prof. Guan Yang, Zhongyuan University of Technology, China  
Prof. Haifeng Sang, Shenyang University of Technology, China  
Prof. Yuantao Chen, Changsha University of Science and Technology, China

**Technical Program Committees**

- Dr. Oluwarotimi Williams Samuel, Chinese Academy of Sciences, China  
Prof. Jian Yao, Wuhan University, China  
Prof. Deyu Zhang, Central South University, China  
Prof. Yuanzhi Wang, Anqing Normal University, China  
Prof. Libor Pekař, Tomas Bata University in Zlín, Czech Republic  
Prof. Hongwei Ge, Dalian University of Technology, China  
Prof. Xiushan Nie, Shandong Jianzhu University, China  
Prof. Guang Sun, Hunan University of Finance and Economics, China  
Prof. Ángel A. San-Blas, Miguel Hernández University of Elche, Spain  
Prof. Guan Yang, Zhongyuan University of Technology, China  
Prof. Haifeng Sang, Shenyang University of Technology, China  
Prof. Changli Li, Hohai University, China  
Dr. Seppo Sirkemaa, University of Turku, Finland  
Prof. Qing Wang, Northwestern Polytechnical University, China  
Prof. Xiaofeng Lu, Xi'an University of Technology, China  
Prof. Changjiang Zhang, Zhejiang Normal University, China  
Dr. Zakwan Jaroucheh, Edinburgh Napier University, UK  
Prof. Jiehui Jiang, Shanghai University, China  
Prof. Huaiqing He, Civil Aviation University of China, China  
Prof. Aiping Qu, University of South China, China  
Prof. Jinping Li, University of Jinan, China  
Simon X. Liao, University of Winnipeg, Canada  
Prof. Wei Wei, Xi'an University of Technology, China  
Prof. Lixin Tan, Hunan Agricultural University, China  
Prof. Tiecheng Song, Chongqing University of Posts and Telecommunications, China  
Dr. Lynette Zhu, Chengdu University of Traditional Chinese Medicine, China  
Prof. Zhigang Liu, Northeast Petroleum University, China

ICIVIS 2021 was broadcast live on Tencent, which was highlighted by speeches given by Prof. Yang Xiao, Prof. En Zhu, Prof. Philippe Fournier-Viger, Prof. Yong Wang and Prof. Jian Yao. Apart from the five keynote speeches, two parallel sessions were held in the conference. The topics of Session 1 are Classification and Recognition & Intelligent System and Control, host by Prof. Heng Li (Central South University). Image and Video Processing & Computer and Machine Vision are the topics of Session 2.

The keynotes presented at this conference were listed as follows.

Keynote speaker	Institution	Topic	Time (min)
Prof. Yang Xiao	University of Alabama, USA	Electricity Theft Detection via Modeling Attackers' Behaviors	40
Prof. En Zhu	National University of Defense Technology, China	Vision feature computing and anomaly detection	40
Prof. Philippe Fournier-Viger	Harbin Institute of Technology, Canadian	Algorithms to discover interesting patterns to improve the design of intelligent systems	40
Prof. Yong Wang	Central South University, China	A Simple Encoding Mechanism in Intelligent Optimization and Its Application to Mobile Edge Computing	40
Prof. Jian Yao	Wuhan University, China	Challenging issues and key technologies for multi-image fusion	40

The committee has received 213 submissions from four regions, then selected 106 peer-reviewed full papers to be published in this proceeding. On behalf of the organizing committee, we also thank the members of the organizing committees and the program committees.

Last but not least, we wish to express our heartfelt appreciation to the keynote speakers, reviewers, editors and academicians for their kind help and support. ICIVIS 2021 thanks all the authors, participants and Springer for their great contribution that made this conference possible and all the hard work worthwhile.

We appreciate your attendance and your share in the ICIVIS 2021.

Wuhan, China  
Tuscaloosa, USA  
Changsha, China  
Changsha, China

Prof. Jian Yao  
Prof. Yang Xiao  
Prof. Peng You  
Prof. Guang Sun

# **Introduction**

This book presents peer-reviewed articles from The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021), held at Changsha, China. It presents original research results, innovative ideas and the current hot issues learnt that touch on many aspects of computer vision, image processing and pattern recognition. ICIVIS is an international conference that serves researchers, scholars, professionals, students and academicians looking to foster both working relationships and gain access to the latest research results. Topics covered include Image Processing & Intelligent Control.

# Contents

## Image and Version

<b>Research on Multi-label of Gastritis Pathological Images Based on Weakly Supervised Deep Learning .....</b>	3
Haoyang Cui, Xie Ding, Jingyi Zhang, Dan Huang, Yi Wang, Qinghua You, Boqiang Zhang, Yu Wang, and Jiaxu Zhao	
<b>A Survey of Object Tracking Methods Based on Deep Learning .....</b>	21
Yang Yi, Zijian Meng, and Guixiong Tian	
<b>Hyperspectral Image Classification Based on Stacked Extreme Learning Machine .....</b>	33
Qiongying Fu, Jinchun Qin, and Li Li	
<b>Nasopharyngeal Organ Segmentation Algorithm Based on Dilated Convolution Feature Pyramid .....</b>	45
Xiaoying Pan, Dong Dai, Hongyu Wang, Xingxing Liu, and Weidong Bai	
<b>A Survey of FPGA-Based Deep Learning Acceleration Research .....</b>	59
Ziyi Lv and Jing Zhang	
<b>A Survey of Track Prediction Method of AUV Based on Deep Learning .....</b>	67
Yuna Yu, Jing Zhang, and Tianchi Zhang	
<b>Recognition of Farmers' Working Based on HC-LSTM Model .....</b>	77
Wenxin Zhao, Jinpo Xu, Xiang Li, Zhaoqi Chen, and Xin Chen	
<b>Recognition of Corn Diseases and Insect Pests Based on Residual Network and Transfer Learning .....</b>	87
Chun Liao, Jiahao Wang, Qilin Xiong, and Wanlin Gao	

<b>A Lightweight Image Super-Resolution Network Based on ESRGAN for Rapid Tomato Leaf Disease Classification .....</b>	97
Lei Zha, Yangjing Shi, and Juan Wen	
<b>Fuzzy Image Processing Based on Deep Learning: A Survey .....</b>	111
Shoucun Chen, Jing Zhang, and Tianchi Zhang	
<b>Semi-supervised Generative Adversarial Network for Face Anti-spoofing .....</b>	121
Junting Chen, Jiwen Dong, Qingtao Hou, Shenyuan Li, Xizhan Gao, and Sijie Niu	
<b>Improving Apple Detection Using RetinaNet .....</b>	131
Zhen Ma and Nianqiang Li	
<b>Megavoltage Computed Tomography (MVCT) Imaging Quality Improvement via Convolutional Neural Network .....</b>	143
Zengjing Zhao, Jiwen Dong, Sijie Niu, Yan Zhang, and Jian Zhu	
<b>Research and Application of Railway Turnout Gap Detection Based on Improved Canny Algorithm .....</b>	151
Xinpeng Liu, Runyuan Sun, and Zhifeng Liang	
<b>3D Vision Transformer for Postoperative Recurrence Risk Prediction of Liver Cancer .....</b>	163
Fan Li, Xueying Zhou, Xizhan Gao, Hui Zhao, and Sijie Niu	
<b>Attention-Aware U-Net Network for Segmentation of Retinopathy Region .....</b>	173
Wenyang Kong, Fan Li, Ruiwen Xing, Xizhan Gao, Hui Zhao, Jie Su, and Sijie Niu	
<b>LESN: Low-Light Image Enhancement via Siamese Network .....</b>	183
Xixi Nie, Zilong Song, Bing Zhou, and Yating Wei	
<b>A Lightweight-Improved CNN Based on VGG16 for Identification and Classification of Rice Diseases and Pests .....</b>	195
Kaibo Liang, Yuzhi Wang, Li Sun, Dongpeng Xin, and ZiWei Chang	
<b>Compressed Channel Attention Mechanism for 3D Medical Image Segmentation of Liver .....</b>	209
Yuwei Liao, Lianglun Cheng, and Weida Lin	
<b>Action Detection Based on Transfer Learning of Human Pose Estimation .....</b>	219
Weida Lin, Zhuowei Wang, and Yuwei Liao	
<b>Fine-Grained Visual Classification Based on Wisely Feature Map Filtering Mechanism .....</b>	229
Haiyuan Chen, Lianglun Cheng, and Ganghan Zhang	

Contents	xiii
<b>Study on Prostate Image Segmentation Using Improved U-NET</b> .....	237
Mengya Sun	
<b>Music Auto-tagging Based on Attention Mechanism and Multi-label Classification</b> .....	245
Chen Ju, Lixin Han, and Guozheng Peng	
<b>Prediction of Apoplexy Syndrome Based on Graph Neural Network</b> .....	257
Shuoyan Zhang, Zhuangzhi Yan, Jiehui Jiang, and Tianyu Gu	
<b>Traditional Chinese Medicine Information Analysis Based on Multi-task Joint Learning Model</b> .....	267
Chenyuan Hu, Zhuangzhi Yan, Jiehui Jiang, Shuoyan Zhang, and Tianyu Gu	
<b>TongueCaps: A Model for the Multiclassification of Tongue Color</b> .....	279
Jinghong Ni, Zhuangzhi Yan, and Jiehui Jiang	
<b>Investigation of Multi-task Learning for Object Detection</b> .....	291
Yujie Zhang, Dongsheng Li, and Junping Xiang	
<b>Design of Place Recognition Algorithm Based on VLAD Code and Convolutional Neural Network</b> .....	297
Bo Wang, Xinsheng Wu, An Chen, and Hongxia Gao	
<b>Fusing Global Gabor Feature and Local Binary Pattern for Texture Image Recognition</b> .....	309
Junmin Wang	
<b>Multi Association Semantics-Based User Matching Algorithm Without Prior Knowledge</b> .....	321
Qiuyan Jiang and Daofu Gong	
<b>Robust Template Matching via Hierarchical Convolutional Features from a Shape Biased CNN</b> .....	333
Bo Gao and Michael W. Spratling	
<b>Reach on Visual Image Restoration Method for AUV Autonomous Operation: A Survey</b> .....	345
Teng Xue, Jing Zhang, and Tianchi Zhang	
<b>On Improving Perceptual Image Hashing Using Reference Image Construction</b> .....	353
Xinran Li and Zichi Wang	
<b>A Zero-Watermarking Against Large-Scale Cropping Attack</b> .....	365
Jing Wang, Sellappan Palaniappan, and Bing He	

<b>A Brief Review of Image Dehazing Algorithms Based on Deep Learning .....</b>	<b>377</b>
Juan Wang, Chang Ding, Minghu Wu, Yuanyuan Liu, and Guanhui Chen	
<b>A Study on the Importance of Tactile Stimulation on Immersive Experience for Digital Communication .....</b>	<b>393</b>
Jia Feng, Zhe Qian, Guoli Chen, and Wei Wang	
<b>Single Shot Tooth Mark Detector for Tongue Diagnosis in Traditional Chinese Medicine .....</b>	<b>403</b>
Xiaodong Huang and Li Zhuo	
<b>Based on Machine Vision, Automatic Measuring System for Adhesive Coating of Caliper Tool .....</b>	<b>413</b>
Xiaofei Wang, Xiaolei Zhang, Jing Wang, and Hua Fan	
<b>Infrared Image Data Augmentation Based on Improved Image-to-Image Translation Network .....</b>	<b>423</b>
Zizhuang Song, Jiawei Yang, Dongfang Zhang, and Yue Zhang	
<b>Rotation Invariant Convolutional Neural Network Based on Orientation Pooling and Covariance Pooling .....</b>	<b>433</b>
Xiaoqin Yao, Tiecheng Song, Jingying Zeng, and Yangming Xie	
<b>Deep Spatial-Temporal Graph Modeling of Urban Traffic Accident Prediction .....</b>	<b>445</b>
Yongxian Huang, Fan Zhang, and Jinhui Hu	
<b>Exposing Video Frame Removal via Deep Features .....</b>	<b>457</b>
Tianle Wu, Chunhui Feng, and Yigong Huang	
<b>Research on Semantic Segmentation and Object Grasping Strategy Generation Based on Deeplab Algorithm .....</b>	<b>467</b>
Shaobo Li, Qiang Bai, Jing Yang, Liya Yu, and Guangwei Wang	
<b>Comparison of SAR Image Water Extraction Algorithms Based on Grey Incidence Analysis .....</b>	<b>477</b>
Jingjue Chen, Rui Liu, Mei Yang, Xin Yang, Yuan Tao Yang, and Tianqiang Liu	
<b>Combine Local and Global Feature Extraction for Point Cloud Classification .....</b>	<b>489</b>
Xiaolong Lu, Baodi Liu, Weifeng Liu, Kai Zhang, Ye Li, and Peng Liu	
<b>Image Detection of Peach Diseases and Pests .....</b>	<b>501</b>
Qi Li, Wenjie Sun, Aiju Shi, Chengmin Lei, and Shaomin Mu	
<b>Automatic Classification of Tongue Shape Based on Improved Analytic Hierarchy Process .....</b>	<b>515</b>
Shanshan Gao, Liqian Zhang, Menghang Li, and Wenhan Dou	

<b>Contents</b>	<b>xv</b>
<b>Design of Visualization System for Digitalization of the Discrete Manufacturing Industry .....</b>	<b>529</b>
Jianguo Yan, Yu Zhao, Wei Wang, Fangfu Xu, Wei Zhu, and Lili Jin	
<b>Research and Simulation of Image Specific Region Recognition Technology .....</b>	<b>539</b>
Nan Li, Chang Jiang Feng, and Bin Lang	
<b>Guided Filter in Least Squares to Remove Non-uniform Strong Noise of Underwater Target Image .....</b>	<b>551</b>
Guang Liu, Shikang Wu, Yu Shi, and Xia Hua	
<b>Optical Flow Fusion Synthesis Based on Adversarial Learning from Videos for Facial Action Unit Detection .....</b>	<b>561</b>
Shuangjiang He, Huijuan Zhao, Jing Juan, Zhe Dong, and Zhi Tao	
<b>A Brief Survey on Privacy-Preserving Methods for Graph-Structured Data .....</b>	<b>573</b>
Yunan Zhang, Tao Wu, Xingping Xian, and Yuqing Xu	
<b>Disentangled Representation Learning from Videos for Facial Action Unit Detection .....</b>	<b>585</b>
Zhe Dong, Huijuan Zhao, Jing Juan, Shuangjiang He, and Zhi Tao	
<b>Facial Expression Recognition Based on Images Captured and Refined with Synchronized Voice Activity Detection .....</b>	<b>597</b>
Xiaoqing Jiang, Lingyin Wang, and Yue Zhao	
<b>An Unsupervised Concrete Crack Detection Method Based on nnU-Net .....</b>	<b>609</b>
Xinyang Li, Shaowu Yang, and Hengzhu Liu	
<b>Robust Facial Landmark Localization Based on Texture and Pose Correlated Initialization .....</b>	<b>625</b>
Junwei Zhou, Mengying Li, and Yiyun Pan	
<b>An Accurate Visual Navigation Method for Wheeled Robot in Unstructured Outdoor Environment Based on Virtual Navigation Line .....</b>	<b>635</b>
Zhen Liang, Tiyu Fang, Zihao Dong, and Jinping Li	
<b>DABU-Net: Dilated Convolution and Attention U-Net with Boundary Augment for Medical Image Segmentation .....</b>	<b>657</b>
Ye Yuan, Yajing An, and Guoqiang Zhong	
<b>Building Boundary Vectorization from Satellite Images Using Generative Adversarial Networks .....</b>	<b>671</b>
Kunyue Yan, Yingxiao Xu, and Hao Chen	

<b>Research on Tomato Maturity Detection Based on Machine Vision</b>	679
Sen Lian, Linlin Li, Weibin Tan, and Lixin Tan	
<b>Correlation Filter RGB-T Tracker with Modality and Channel Reliability</b>	691
Fei Zhang and Shiping Ma	
<b>Local Binary Complement Pattern for Color-Inversion Invariant Texture Classification</b>	703
Yuqian Wu and Tiecheng Song	
<b>Video Instance Segmentation of Rock Particle Based on MaskTrack R-CNN</b>	715
Man Chen, Maojun Li, and Yiwei Li	
<b>Infrared Dim Target Detection Based on Convolutional Neural Networks</b>	725
Pinghuang Zhou and Wei Ai	
<b>Improvement of Multi Frequency Heterodyne Phase Unwrapping in Extreme Environment</b>	733
Bingwei Zhang, Junyi Lin, Shaoning Lin, Yabin Liu, and Kaiyong Jiang	
<b>A Quick and Accurate Method to Identify Betel Nut Based on Mobilenetv3</b>	745
Yun Dai, Ming Lu, and Zuguo Chen	
<b>PRM: Pose Recalibration Module for Action Recognition</b>	757
Guixiong Tian, Yang Yi, Zijian Meng, Zhonghong Li, and Jialun Song	
<b>Stereo Visual SLAM System Reasonably Use Point and Line Features</b>	767
Huiyue Qiao, Xuhu Ren, Luyan Niu, Yang Feng, and Songzho Liu	
<b>An Intelligent Foreign Substance Inspection Method for Injection Based on Machine Vision</b>	781
Bowen Zhou, Liang Chen, and Lianghong Wu	
<b>A Modified SiamRPN for Visual Tracking</b>	795
Wei Zhou, Yuxiang Liu, Haixia Xu, and Zhihai Hu	
<b>Unsupervised Person Re-identification via Multi-branch Network</b>	807
Xiaobin Wang, Baodi Liu, and Weifeng Liu	
<b>Affine Non-negative Hybrid Collaborative Representation Based Classification</b>	819
Haoquan Guan, Baodi Liu, Weifeng Liu, Kai Zhang, Ye Li, and Peng Liu	

<b>Pathologist-Level Classification of Melanoma Disease Pathologies Using a Convolutional Neural Network: A Retrospective Study of Chinese</b>	833
Tao Li, Fangfang Li, Jie liu, and Ke Zuo	
<b>Handwritten Digits Recognition Based on Water Drop Algorithm and CNN</b>	841
Geying Liang, Han Long, and Baoliang Dong	
<b>A New Approach Based on Crater Detection and Matching for Self-Localization During Lunar Landings</b>	849
Zhouyuan Qian, Hao Cheng, Tao Hu, Tao Cao, Yu Han, and Liang He	
<b>Intelligent Systems</b>	
<b>Robust Spectral Clustering via the Ordering Metric</b>	863
Bingjie Li, Tianhao Ni, and Zhenyue Zhang	
<b>Influence of Initialization and Modularization on the Performance of Network Morphism-Based Neural Architecture Search</b>	875
Xuehui Chen, Xin Niu, Jingfei Jiang, Hengyue Pan, Peijie Dong, and Zimian Wei	
<b>A Document Image Quality Assessment Method Based on Feature Fusion</b>	889
Weisheng Wang, Zhiyang Yan, and Hongli Lin	
<b>Design of Simulation Device for Greenhouse Control</b>	901
Yunsong Jia, Shuaiqi Huang, Liang Xiao, Shaochen Yang, and Xiang Li	
<b>Vis–NIR Hyperspectral Dimensionality Reduction for Nondestructive Identification of China Northeast Rice</b>	913
Jiahao Wang, Chun Liao, Jingyi Zhao, and Wanlin Gao	
<b>An Encryption Scheme for Internet of Things Monitoring System</b>	923
Haoyi Sun, Shuihai Zhang, Chunli Lv, and Bei Pei	
<b>An Overview of Text Steganalysis</b>	933
Yu Yang, Lei Zha, Ziwei Zhang, and Juan Wen	
<b>Design and Experiment of UAV Variable Spray Control System Based on RBF-PID</b>	945
Yunling Liu, Yan Ma, Bowen Wu, and Yajia Liu	
<b>Research on a Safe and Reliable Agricultural Product Traceability System Driven by Permissioned BlockChain Technology</b>	955
Guofeng Zhang, Xiao Chen, Bin Feng, and Juan Wen	

<b>Simultaneously Learning Syntactic Dependency and Semantics Reasonability for Relation Extraction .....</b>	967
Xin Wang, Nan Yin, Xiang Zhang, Xinyi Bai, and Zhigang Luo	
<b>Overview on Job Running Times Prediction Algorithms for HPC Platform .....</b>	981
Hao Wang and Yiqin Dai	
<b>Effects of Game Perspectives Differences on Immersion Using Eye Tracking .....</b>	993
Peng Li, Xu Jiang, and Xuebai Zhang	
<b>A Tolerance Classes Partition-Based Re-Definition of the Rough Approximations for Incomplete Information System .....</b>	1003
Lei Wang, Bin Liu, Xiangxiang Cai, and Chong Wang	
<b>A Cyber Security Situational Awareness Extraction Method Oriented to Imbalanced Samples .....</b>	1013
Kun Yin, Yu Yang, and Chengpeng Yao	
<b>Mine Cable Fault Distance Detection .....</b>	1027
Zezhong Liu, Ming Lu, Zuguo Chen, Wang Cheng, and Jinyu Wang	
<b>PCNetOP: Partial Completion Network with Order Prediction .....</b>	1037
Yifan Wang and Yongping Xie	
<b>Fastener Identification Method Based on Two-Stage Positioning .....</b>	1047
Yan Li, Hongbin Liu, and Zhigang Liu	
<b>EmbedLOF: A Network Embedding Based Intrusion Detection Method for Organized Attacks .....</b>	1059
Peng Chen, Yunfei Guo, Jianpeng Zhang, and Hongchao Hu	
<b>An IoT Data Transmission Model Based on Push Mechanism .....</b>	1075
Deguo Yang, Zeming Wang, Shuai Qi, and Lianqiang Niu	
<b>Electrical Load Forecasting Using Hybrid of Extreme Gradient Boosting and Light Gradient Boosting Machine .....</b>	1083
Eric Nziyumva, Rong Hu, Chih-Yu Hsu, and Jovial Niyogisubizo	
<b>Research on Splicing of Horizontal and Longitudinal Shredded Paper Based on Hungarian Algorithm .....</b>	1095
Lizhi Shen, Zhixiong He, Lang Chen, and Rongyuan Chen	
<b>Rendezvous Control for Autonomous Underwater Vehicles with Event Triggered Cloud Access .....</b>	1107
Feng Zhou, Ge Zheng, and Kewu Tao	
<b>Towards Distractibility Induced trust Management Using BlockChain for Edge Computing .....</b>	1121
Haochen Yang, Guanghui Wang, Lifeng Dong, and Xin He	

<b>Social Robot Navigation Based on a 2D Gauss-Gumbel Spatial Density Model in Human-Populated Environments .....</b>	1133
Jianfang Lian, Wentao Yu, Kui Xiao, Feng Qu, and Chaofan Liu	
<b>Fool a Hashing-Based Video Retrieval System by Perturbing the Last 8 Frames of a Video .....</b>	1145
Chao Hu, Liang Huang, and Ronghua Shi	
<b>Multi-level Road Damage Identification Algorithm Based on Vehicle-Mounted Smartphone .....</b>	1155
Deng Ma, Kai Gao, and Ronghua Du	
<b>A Digital Twin Model for Battery Management Systems: Concepts, Algorithms, and Platforms .....</b>	1165
Mi Zhou, Lu Bai, Jiaxuan Lei, Yibin Wang, and Heng Li	
<b>A Research on Remaining Useful Life of Solenoid Valve Based on Millimeter Wave Radar .....</b>	1177
Xin Liu, Shou Li, Weirong Liu, and Feng Zhou	
<b>A Multi-link Data Congestion Control Algorithm in Spatial Delay Tolerance Network .....</b>	1185
Li Yi and Renjie Zhang	
<b>Data-Driven Fault Prognosis for Pneumatic Valves in Train Electropneumatic Brake System .....</b>	1195
Dianzhu Gao, Jun Peng, Ning Ding, and Yingze Yang	
<b>A RCS Periodicity Extraction Algorithm for Ballistic Target .....</b>	1207
Chaowei Li, Bing Xie, and Yu Pei	

## **Image and Version**

# Research on Multi-label of Gastritis Pathological Images Based on Weakly Supervised Deep Learning



Haoyang Cui · Xie Ding · Jingyi Zhang · Dan Huang · Yi Wang ·  
Qinghua You · Boqiang Zhang · Yu Wang · and Jiaxu Zhao

**Abstract** Computer-aided diagnosis technology based on artificial intelligence has been widely used in the medical field, especially in pathological image diagnosis (Whole Slide Image, WSI). This paper mainly introduces the attention mechanism-based multiple instance networks (Attention-MIL) to detect three clinical indicators of “activity”, “atrophy” and “intestinal metaplasia” in gastritis. The model uses multiple parallel attention branches to automatically identify multiple tissue regions with high diagnostic value, and aggregate them with their corresponding attention scores to form WSI-level features. According to the gastritis pathology report, we assign three labels to each WSI. In the independent test analysis, comprehensive evaluation has fully proved the effectiveness and feasibility of this method in the multi-label prediction task of gastritis. Besides, the Attention-MIL model can visualize highly concerned lesion areas, which to a certain extent provides interpretability for clinical research. The method in this paper reveals the great potential of using weakly supervised learning to achieve assisted diagnosis of gastritis.

---

H. Cui · X. Ding (✉) · J. Zhang · Y. Wang · J. Zhao  
The Data Center, Wonders Information Co., LTD., 1518 Lian'hang Road, Shanghai 201112, China  
e-mail: [chyljh@shu.edu.cn](mailto:chyljh@shu.edu.cn)

D. Huang  
Department of Pathology, Fudan University Shanghai Cancer Center, 270 Dong'an Road,  
Shanghai 200032, China

Y. Wang  
Department of Information Center, Fudan University Shanghai Cancer Center, 270 Dong'an  
Road, Shanghai 200032, China

Shanghai Engineering Research Center of Artificial Intelligence Technology for Tumor Diseases,  
270 Dong'an Road, Shanghai 200032, China

Q. You  
Departments of Pathology, Shanghai Pudong Hospital, Fudan University Pudong Medical Center,  
2800 Gong'wei Road, Shanghai 201399, China

B. Zhang  
Shanghai Foremost Medical Technology Co., LTD., 258 Lian'chuan Road, Shanghai 201112,  
China

**Keywords** Pathological image · Weakly supervised learning · Attention mechanism · Multi-label

## 1 Introduction

Chronic gastritis is a common disease of the digestive system, and it is also a pre-neoplastic lesion of gastric cancer [1]. Pathological biopsy is very important for the diagnosis of chronic gastritis. Relevant research on gastric mucosal atrophy and metaplasia show that the clinical pathological evaluation of chronic gastritis mainly includes “activity”, “intestinal metaplasia” and “atrophy”. Standardized and unified pathology reports provide clinicians with sufficient information. However, the morphological evaluation of multiple pathological indicators increases the workload of the pathologist. With the maturity and standardization of gastroscopy biopsy technology, the number of gastroscopy biopsy samples is continuously increasing, but there is a serious shortage of pathologists in China. In addition, the diagnosis results will be affected by factors such as fatigue, subjective experience, etc. In recent years, the rapid development of digital pathology scanning technology [2] and artificial intelligence (AI) has provided great potential for pathomorphological diagnosis. Through deep learning, optimized models and algorithms, AI can continuously approach the level of clinical diagnosis, avoid misdiagnosis due to fatigue and subjective factors, and assist pathologists in completing pathological diagnosis efficiently and accurately [3].

## 2 Background Work

The outstanding performance of deep learning in computer vision, such as image classification, semantic segmentation and object detection [4–6]. At present, automatic recognition technology based on deep learning has also achieved certain results in the diagnosis of digital pathological images, such as breast cancer, lung cancer, bowel cancer and prostate cancer [7–10]. Most of these studies use fully supervised learning, and complete training by manually annotating pixel-level tumor cells, and improving the accuracy of model recognition through feedforward and feedback operations. However, processing WSI with a scale of hundreds of millions of pixels is the main challenge of fully supervised learning. A single pathological image obtained at  $20\times$  magnification can contain billions of pixels, and the area we need to focus on may only be thousands of pixels, so a large number of experienced pathologists are required to manually label the tumor area [11]. In reality, there is a lack of expert-level pathologists, and the reports corresponding to pathological slices can only provide category information. Therefore, deep learning models based on weak supervision have begun to be applied to pathological image-assisted diagnosis. Weakly supervised learning is suitable for situations that are greatly affected by subjective factors

or it is difficult to obtain annotations of tumor regions. For non-tumor diseases, it is often manifested as abnormal cell proportions or abnormal distribution. Evaluation is highly subjective, and it is difficult to directly Annotation tumor cells. Weakly supervised learning can give full play to its advantages. Recently, a remarkable work by Campanella et al. [7]. We use the RNN model to integrate semantically rich feature representations across patch-level instances to obtain the final slide-level diagnosis. In their method, the authors successfully obtained an AUC greater than 0.98, which can detect four types of cancer in an extensive multi-center data set (44,732 WSI) without the need for expensive pixel-by-pixel manual annotation. For gastric cancer, Sharma et al. [12] used a dataset of 11 WSIs to perform carcinoma classification and proposed a CNN architecture using WSI automatic classification in histopathology, thus revealing the practicality of artificial intelligence in the research of gastric cancer digital pathology. But most weakly supervised models are mainly for binary classification problems and are not suitable for multi-label problems. Therefore, in order to adapt to the clinical environment more widely, realize the automatic identification of multiple pathological indicators related to gastritis.

In this paper, attention mechanism based multiple instance network (Attention-MIL) is proposed to solve the problem of weakly supervised classification of gastritis pathological images. This method introduces the attention mechanism [13, 14] into multi-instance learning, and uses  $n$  parallel attention branches to focus on different tissues forms in the same WSI. Generate  $n$  different WSI-level features in a similar Multi-Tasking [15, 16] manner, where each feature is determined by a different tissue concerned by the network. Construct  $n$  independent classifiers at the same time to obtain the attention score of each category corresponding to the WSI level, and finally predict the probability of each category of with Sigmoid function.

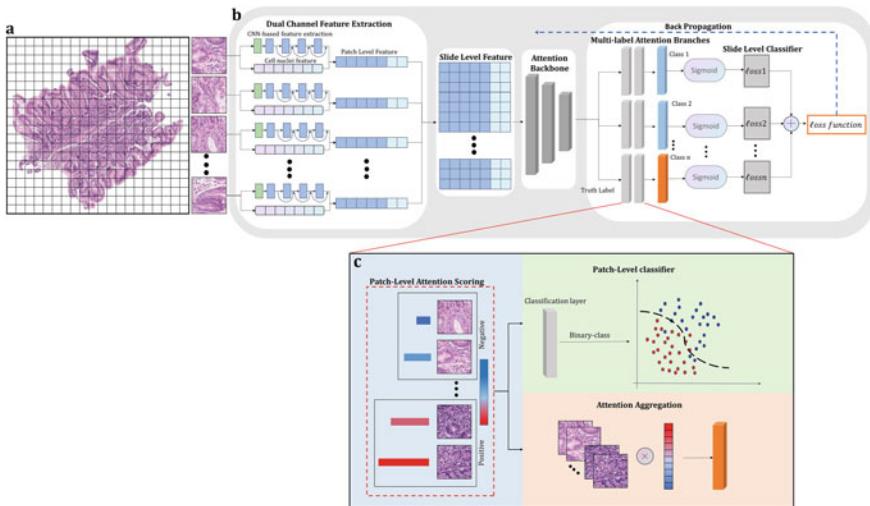
### 3 Methods

The weakly supervised learning framework proposed in this paper is used to solve the multi-label problem of gastritis pathological images. The required gastritis data is only a single sample of known WSI-level diagnosis results, and there is no pixel-level regional annotation in any area of WSI. Attention-MIL is built on the framework of Multi-Instance [17] and belongs to the category of weakly supervised learning. The framework treats each WSI as a package of thousands to tens of thousands of smaller patches, and these patches are treated as instance. The original MIL [18] algorithm is mainly for the binary classification problem of positive and negative. It assumes that at least one block in a positive WSI is a positive, and all blocks in a negative WSI are negative. Its aggregation function is a form of max pooling. It simply takes the block with the largest probability as the prediction result. On the one hand, it makes the prediction error accumulate. On the other hand, it is not suitable for the multi-label problem of gastritis. The Attention-MIL method proposed in this article introduces the attention mechanism into multi-instance learning and uses  $n$  parallel attention branches to generate different feature representations of the same

WSI. Each feature representation is determined by the attention mechanism from different tissue features in the WSI image, and these tissue features are regarded as the most representative feature representation of one of the n categories in the multi-label diagnostic task. And after each attention branch, an independent classifier is constructed to generate the attention score corresponding to the WSI level of each category. Finally, predict the category probability through Sigmoid.

### 3.1 Multi-instance Network Based on Attention Mechanism (Attention-Mil)

The Attention-MIL framework proposed in this paper is shown in Fig. 1. From Fig. 1b, it can be seen that Attention-MIL does not directly train the patches, but uses a dual-channel feature extraction module to extract features for each patch. As shown in Fig. 1a, the whole WSI is segmented and input into the feature extraction module in Fig. 1b. One channel is feature extraction based on pre-trained resnet50, and the other channel is a histopathological feature based on manual extraction. After



**Fig. 1** Overview of the Attention-MIL conceptual framework. **a** Following segmentation, image patches are extracted from the tissue regions of the WSI. **b** CNN encoding and artificial feature encoding are used to form the feature vector of the image patches. During training and inference, the feature vector of each image patches is passed to Attention-MIL. **c** For each WSI, the attention mechanism will assign attention scores according to the importance of each image patches to WSI diagnosis, and aggregate it with the features of the image patches to form a WSI-level feature representation for the final classification prediction. Use the features of strong and weak attention regions as representative samples to train a binary classifier to learn the differences between the features of different image patches to distinguish between positive samples and negative samples

that, the two sets of features are spliced to form a 1024-dimensional feature vector. Therefore, the 1024-dimensional vector corresponding to each patch is used as the training data of Attention-MIL model.

In the design of the attention mechanism, as shown in Fig. 1b, the prediction result of the attention network is corresponding to the number of categories  $n$  to form  $n$  parallel attention branches. Each branch will highly focus on different tissue areas in WSI, and assign an attention score to each patch in the focus area. Aggregate the features of all patches and their corresponding attention scores to form a WSI-level feature representation. For the same WSI,  $n$  different WSI-level feature representations can be obtained, which can be used for WSI-level multi-label prediction. Therefore, each attention branch can learn its corresponding category, to determine which regions have a large contribution to the WSI prediction result and assign corresponding attention scores. In Attention-MIL, the first fully connected layer  $W_1 \in \mathbb{R}^{512 \times 1024}$  further compresses each fixed 1024-dimensional patch-level representation  $z_k$  to a 512-dimensional vector  $h_k = W_1 z_k^T \in \mathbb{R}^{512 \times 1}$ . We use the first three layers of the attention network as the backbone of the shared attention, and then splits the attention network into  $n$  parallel attention branches. In each attention branch, introduce a trainable linear classifier for patch-level classification.

As shown in Fig. 1c, in each attention branch, the attention score of each patch and the corresponding feature vector can be obtained. A patch with a high attention score is regarded as a strong attention area, and a patch with a low score is regarded as weak attention area. We use strong and weak attention patches as training samples to train a linear classifier that can distinguish between positive and negative patches. Map the attention score of each patch to the original WSI and present it in the form of a heat map, so that the suspected lesion area can be visualized, making Attention-MIL interpretable in clinical research.

**Patch Feature Extraction.** For each digitized slide, first segmentation the tissue area. After segmentation, for each slide, our algorithm crops  $256 \times 256$  patches from the segmented foreground contours at the 20 magnification. After cropping, two sets of feature encoding are performed respectively: One set of pre-trained *ResNet50* [19] for low-dimensional coding of patches. The other set uses *Hover-Net* [20, 21] to segment cells to obtain cell shape and appearance characteristics. Here, we use 17 representative feature descriptions: mean nuclei intensity; average fore-/background difference; standard deviation of nuclei intensity; Gray Level Co-occurrence Matrix (GLCM) of dissimilarity; GLCM of homogeneity; GLCM of energy; GLCM of Angular Second Moment (ASM); eccentricity; area, maximum length of axis; minimum length of axis; perimeter, solidity; orientation and centroid coordinates. Then the two sets of encoded of features are spliced to form a 1024-dimensional feature vector to represent each patch. Therefore, this article uses the extracted 1024-dimensional features as the training data of the Attention-MIL network. Compared with the original pixel value, we can put the feature vectors corresponding to all the patches of a WSI segmentation into the GPU memory at the same time, thereby avoiding sampling the patches. This not only speeds up the training time but also reduces the computational cost.

**Attention Mechanism.** In the design of multi-category attention mechanism,  $n$  categories correspond to  $n$  different attention branches. The feature map output by the first fully connected layer is  $I \in \mathbb{R}^{N \times 512}$ , each row in the feature map represents a patch feature in WSI. The structure diagram of the attention mechanism is shown in Fig. 2. The weights of the three layers are  $U_a \in \mathbb{R}^{512 \times 256}$ ,  $V_b \in \mathbb{R}^{512 \times 256}$  and  $Z_c \in \mathbb{R}^{512 \times 256}$ , which are viewed as the backbone of shared attention. As shown in Fig. 2a,  $g(x)$ ,  $h(x)$  and  $f(x)$  are defined as ( $\otimes$  : Matrix product,  $\odot$  : Hadamard product):

$$g(x) = \text{sigm}(I \otimes U_a) \in \mathbb{R}^{N \times 256} \quad (1)$$

$$h(x) = \tanh(I \otimes V_b) \in \mathbb{R}^{N \times 256} \quad (2)$$

$$f(x) = \text{ReLU}(I \otimes Z_c) \in \mathbb{R}^{N \times 256} \quad (3)$$

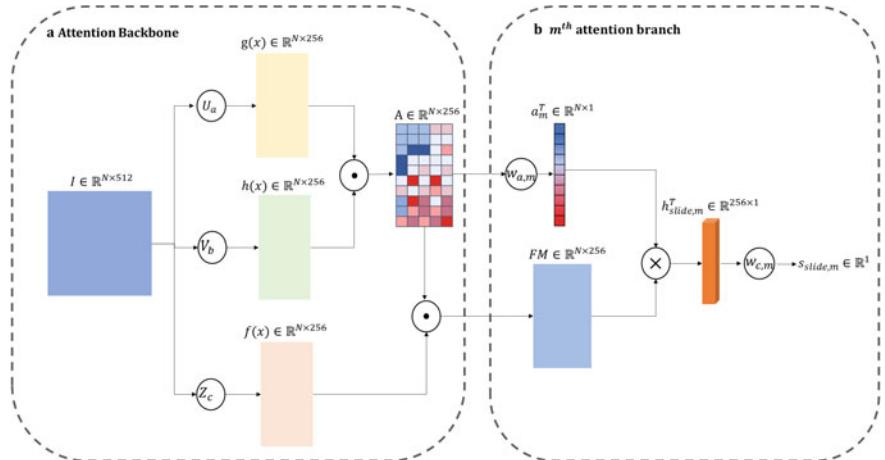
The Attention Map is defined as:

$$A = (g(x) \odot h(x)) \in \mathbb{R}^{N \times 256} \quad (4)$$

Then Attention Feature Map is defined as:

$$FM = (A \odot f(x)) \in \mathbb{R}^{N \times 256} \quad (5)$$

Then, the attention network is divided into  $n$  parallel attention branches  $W_{a,1}, \dots, W_{a,n} \in \mathbb{R}^{1 \times 256}$ . Similarly,  $n$  parallel independent classifiers  $W_{c,1}, \dots, W_{c,n} \in \mathbb{R}^{1 \times 256}$  are constructed to score each category-specific slide level,



**Fig. 2** Overview diagram of the attention mechanism. **a** Block diagram of shared attention backbone network structure. **b** Block diagram of parallel attention branch structure

as shown in Fig. 2b. Therefore, the attention score of the  $k$ th patch for the  $m$ th category is expressed as  $a_{k,m}$ , which is given by Eq. (6), and the slide-level is expressed as aggregation according to the attention score distribution of the  $m$ th category, which is expressed as  $h_{slide,m} \in \mathbb{R}^{1 \times 256}$ , is given by Eq. (7).

$$a_{k,m} = \frac{\exp\{W_{a,m}(\tanh(h_k V_b)^T \odot \text{sigm}(h_k U_a)^T)\}}{\sum_{j=1}^N \exp\{W_{a,m}(\tanh(h_j V_b)^T \odot \text{sigm}(h_j U_a)^T)\}} \quad (6)$$

$$h_{slide,m} = a_m \otimes FM \quad (7)$$

**Slide Diagnosis and Instance-Level Classification.**  $N$  parallel, independent classifiers,  $W_{c,1}, \dots, W_{c,n} \in \mathbb{R}^{1 \times 256}$  are built to score different class-specific of the same slide-level, as shown in Fig. 2b. The corresponding unnormalized slide-level score  $s_{slide,m}$  is given via the classifier layer by  $s_{slide,m} = W_{c,m} h_{slide,m}^T$ . By applying the Sigmoid function to the slide-level prediction scores, the prediction probability distribution for each category can be calculated. Therefore, each attention branch of Attention-MIL can be regarded as a judgment of a specific category. For example, the 1th attention branch is to determine whether the WSI contains “Activity” indicators for gastritis; the 2th attention branch is to determine whether the WSI contains “Intestinal Metaplasia” indicators for gastritis; the 3th attention branch is to determine whether the WSI contains “Atrophy” indicators for gastritis.

In order to obtain highly suspected tissue regions,  $K$  patches with high attention scores are used as positive samples, and  $K$  patches with low attention are used as negative samples to train a binary-class classifier. For each of the  $n$  categories, we place a classification layer with 512 hidden units in each branch of interest. If we express the weights of the classification network that corresponds to the  $m$ th class as  $W_{inst,m} \in \mathbb{R}^{2 \times 512}$ , the assignment scores predicted for the  $k$ th patch, denoted by  $P_{m,k}$  is given as:

$$P_{m,k} = W_{inst,m} h_k \quad (8)$$

In the case of a given WSI true label  $Y$ , assuming that a certain WSI category is not  $Y$ , then all patches in WSI do not belong to  $Y$ , so the  $K$  patches with the highest attention scores will be considered as false positives. Therefore, the binary classifier of training patches helps the network learn the difference between classes.

Training details of Attention-MIL. The total loss of a given slide  $\ell_{total}$  is the sum of the slide-level classification  $\ell_{slide}$  and the instance-level loss  $\ell_{patch}$ , and can be optionally scaled by the scalars  $c_1$  and  $c_2$ .

$$\ell_{total} = c_1 \ell_{slide} + c_2 \ell_{patch} \quad (9)$$

To compute  $\ell_{slide}$  using Binary Cross Entropy With Logits Loss as the loss function and to compute  $\ell_{patch}$  using binary-class SVM [22] as the loss function. We used K

= 50 and weights  $c_1 = 0.7$ ,  $c_2 = 0.3$ . The model optimization function uses Adam [23] optimizer, the weight of L2 decays is  $1e-5$ , and the learning rate is  $2e-4$ .

### 3.2 Visualization

Combine the prediction results of the model with clinical research to provide strong evidence for the interpretability of the model. According to the importance of the patch to the WSI prediction results, the attention mechanism will assign corresponding attention scores to the patch. In the inference process, we use the attention branch corresponding to the predicted category of the model to calculate and save the non-standardized attention score of all patches extracted from the slide. These attention scores are converted to percentiles and scaled from 0 to 1. The threshold value is 0.5 (positive probability is greater than 0.5, negative probability is less than 0.5), and displayed on their respective spatial positions on the slide to visually identify and interpret areas of high attention shown in red (positive evidence, high contribution to model's prediction relatively to other patches) and low-interest areas shown in blue (negative evidence, relative to other patches that contribute less to the model's prediction). The heat map covers the original WSI with a transparency value of 0.4

For each slide in the independent test queue, in addition to the 512-dimensional feature representation after the first fully connected layer, we also record the binary classification probability prediction made by each of the n classification branches. We use PCA to reduce each patch-level feature vector to two dimensions, and then give the feature distribution of positive and negative patches.

## 4 Results and Discussion

### 4.1 WSI Datasets

The dataset comes from 552 cases of biopsy pathology diagnosed with gastritis in the Cancer Hospital of Fudan University (abbreviation: FZ) in 2018. Due to different organizational standards and protocols for tissue processing, slide preparation and digitization, the appearance of WSI images may vary greatly. Therefore, it is important to verify whether the model trained under the Attention-MIL weak supervision framework is robust to data source-specific variables. We collected a total of 130 pathological images of gastritis at the Shanghai Pudong Hospital (abbreviation: PD) as independent test cohorts for evaluating the generalization performance of our trained models.

All pathological sections are digitally processed by the scanner and converted to WSI. The WSI label is given by the gastritis pathology report, which contains three

**Table 1** The number distribution of WSI in gastritis (Positive: P, Negative: N)

Label			Train		Test	
Activity	IM	Atrophy			FZ	PD
P	N	N	40		10	10
N	P	N	26		7	7
N	N	P	55		35	35
P	P	N	214		40	50
P	N	P	69		12	15
N	P	P	22		10	10
P	P	P	9		3	3
N	N	N	0		0	0
Total			435		117	130

**Table 2** Distribution of labels number of gastritis activity, intestinal metaplasia, and atrophy (Positive: P, Negative: N)

Class	Train			Test					
				FZ			PD		
	P	N	Total	P	N	Total	P	N	Total
Activity	332	103	435	65	52	117	78	52	130
IM	272	163	435	60	57	117	70	60	130
Atrophy	155	280	435	60	57	117	63	67	130

pathological indicators: “activity”, “atrophy” and “intestinal metaplasia” (abbreviation: IM). The distribution of gastritis WSI is shown in Table 1. These WSIs are divided into training sets and test sets. The training set uses data from the Cancer Hospital of Fudan University, and the test set uses data from the Cancer Hospital of Fudan University and Shanghai Pudong Hospital. In the final evaluation, we evaluate the model performance on an independent test set. Because each WSI may contain multiple labels, the labels of “activity”, “atrophy” and “intestinal metaplasia” are summarized to obtain the distribution of label data, as shown in Table 2.

## 4.2 Performance Measures

In order to comprehensively evaluate the performance of this model in the multi-label task of gastritis pathological images, this paper uses the area under the ROC curve (AUC) to evaluate the performance of the model [24]. By calculating the confusion matrix of the model on the test set, four evaluation indicators on the test set are obtained: Accuracy, Precision, Recall and F1-score. The formula are as follows:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (10)$$

$$\text{Precision} = TP / (TP + FP) \quad (11)$$

$$\text{Recall} = TP / (TP + FN) \quad (12)$$

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (13)$$

where TP is True Positive, TN is Ture Negative, FP is False Positive and FN is False Negative.

### 4.3 Experimental Results

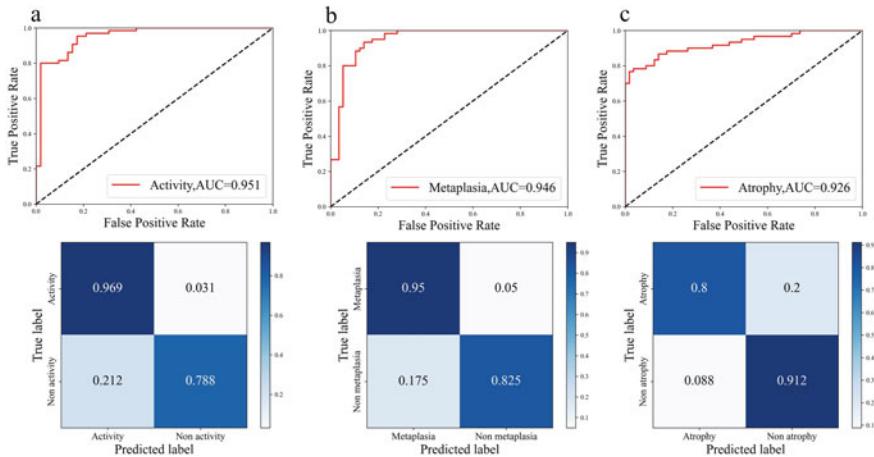
We used multiple hard drives to store the digitized original file of the entire slideshow. Our entire slide processing is implemented in Python and takes advantage of image processing libraries (such as Openslide, Opencv and Pillow). We use the Pytorch deep learning library to build the entire Attention-MIL network. All figures are generated using matplotlib and scikit-learn.

The evaluation results of the Cancer Hospital of Fudan University test data are shown in Table 3, the receiver operating characteristic curve (ROC) for the slide-level classification as well as the corresponding confusion matrix is shown in Fig. 3. From the evaluation results, the recall rates of “activity” and “intestinal metaplasia” both reached 0.95. The AUC value of the three categories are 0.951, 0.946, 0.926. Although the recall rate of atrophy is about 0.8, its AUC value is 0.926, indicating that the model has a certain ability to recognize the clinical features of atrophy. Overall, we noticed the surprising data efficiency of Attention-MIL because it can achieve  $\text{AUC} > 0.92$  tests using only a few hundred slides. Therefore, Attention-MIL is effective and feasible for multi-label prediction of WSI-level of gastritis.

The evaluation results on the Shanghai Pudong Hospital dataset are shown in Table 4. The receiver operating characteristic curve (ROC) curve for the slide-level classification as well as the corresponding confusion matrix is shown in Fig. 4. From

**Table 3** Evaluation results of gastritis Activity, IM, and Atrophy (Test data: FZ)

	Activity	IM	Atrophy	Mean
Accuracy	0.889	0.889	0.855	0.878
Precision	0.851	0.850	0.906	0.869
Recall	0.969	0.950	0.800	0.906
F1-score	0.906	0.899	0.850	0.885
AUC	0.951	0.946	0.926	0.941



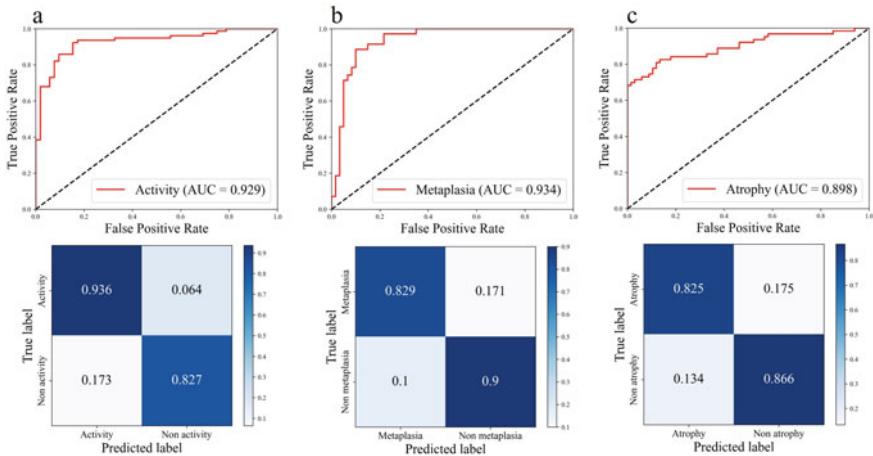
**Fig. 3** ROC curve and confusion matrix on the test data of the Cancer Hospital of Fudan University. **a** ROC curve and normalized confusion matrix for gastritis activity. **b** ROC curve and normalized confusion matrix for gastritis intestinal metaplasia. **c** ROC curve and normalized confusion matrix for gastritis atrophy

**Table 4** Evaluation results of gastritis Activity, IM, and Atrophy (Test data: PD)

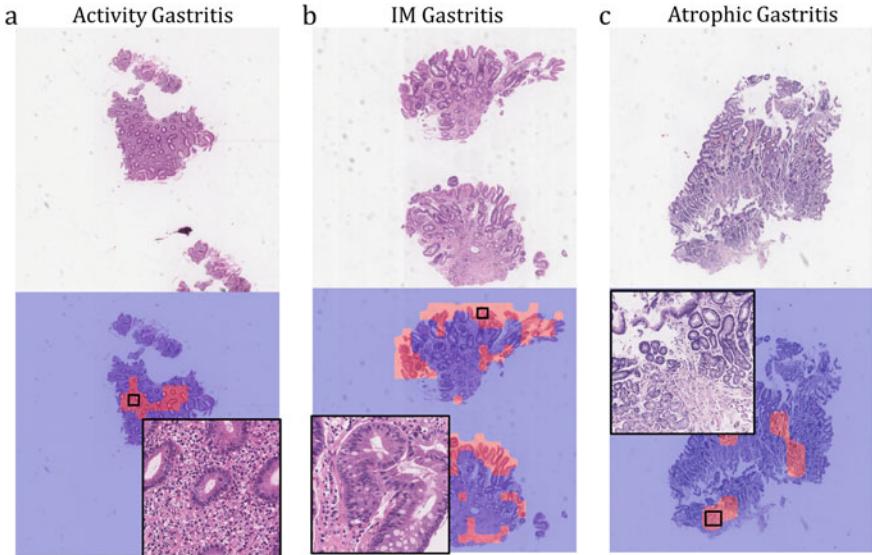
	Activity	IM	Atrophy	Mean
Accuracy	0.892	0.862	0.846	0.866
Precision	0.890	0.906	0.852	0.883
Recall	0.936	0.829	0.825	0.863
F1-score	0.907	0.866	0.839	0.871
AUC	0.929	0.934	0.898	0.920

the evaluation results, the average accuracy of the three categories is 0.866. The AUC values of “activity” and “intestinal metaplasia” are 0.929 and 0.934 respectively. Although the “atrophy” training data is relatively small, its AUC value is still 0.898. The average AUC value of the three categories is 0.92. From the overall evaluation results, it is verified that the model has good generalization performance for data between different institutions.

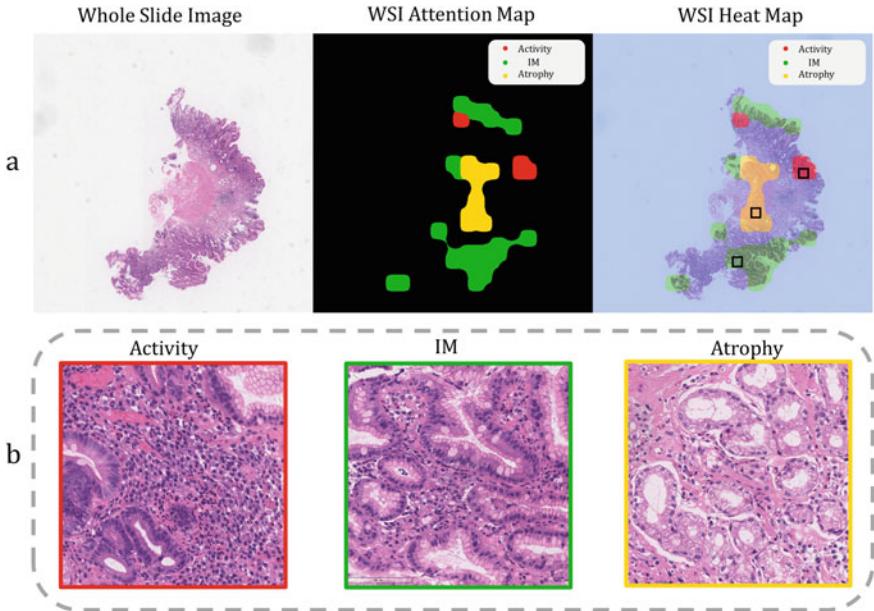
Attention-MIL will assign an attention score to each patch, representing its contribution to the final WSI diagnosis result. Therefore, the prediction results of the model can be combined with the clinical manifestations, and the suspected lesion tissue region can be visualized in the form of a heatmap. Example heatmap output from Attention-MIL model for chronic gastritis using a stride of  $256 \times 256$ . Figure 5 shows the prediction result of only a single lesion, interpret regions of high attention displayed in red and low attention displayed in blue. Figure 6 shows the prediction results of mixed lesions, from which it can be seen that the model can clearly distinguish different types of lesion areas. Therefore, with this simple, intuitive but



**Fig. 4** ROC curve and confusion matrix on the test data of Shanghai Pudong Hospital. **a** ROC curve and normalized confusion matrix for gastritis activity. **b** ROC curve and normalized confusion matrix for gastritis intestinal metaplasia. **c** ROC curve and normalized confusion matrix for gastritis atrophy



**Fig. 5** For the types of lesions in chronic gastritis (**a, b, c**), pathological image of chronic gastritis (top), roughly highlighting the lesions tissue regions (bottom)



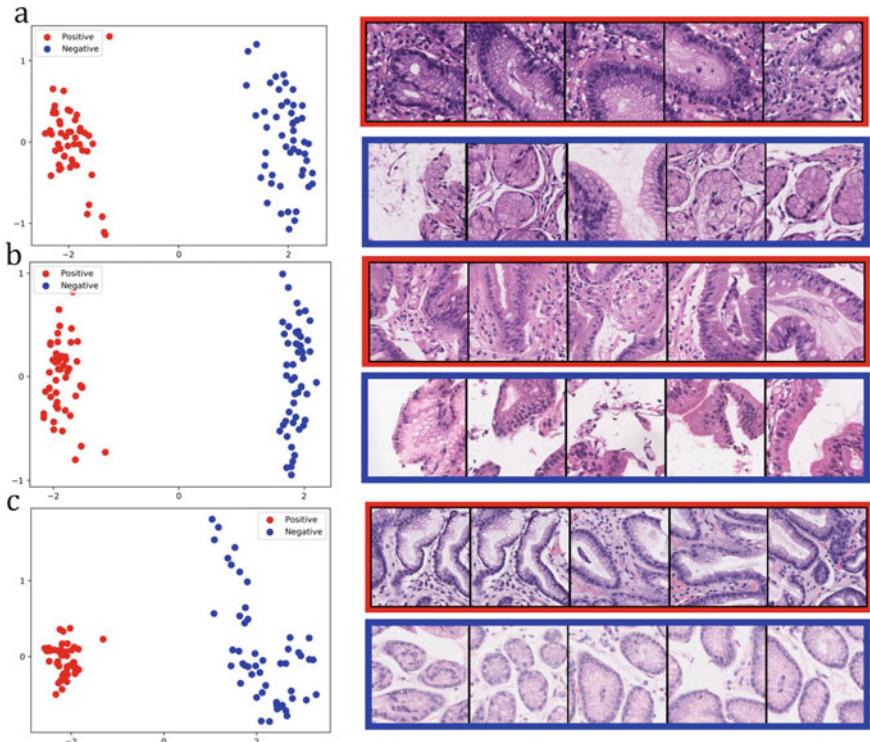
**Fig. 6** Visualize tissue areas of mixed lesions. **a** Roughly highlighting the lesion tissue regions are predicted based on Attention-MIL, Red represents: Activity, Green represents: IM, Yellow represents: Atrophy. **b** Roughly highlight the activity, IM, and atrophy of the lesion tissue area of chronic gastritis

interpretable and visual technology, it can assist pathologists in diagnosis. However, care should be taken not to rely too much on paying attention to heat maps, lest they can be used as pixel-perfect segmentation masks. Intuitively, the attention score of each area in the slide is relative, and only represents the model's interpretation of which areas are more important (relative to other areas) in determining the slide level prediction.

According to the Instance-Level Classification Method, visualize the feature distribution of patches with large contributions and patches with small contributions. In this paper, PCA is used to compress the patches features to two dimensions, thereby visualizing the two-dimensional feature distribution of the patches, as shown in Fig. 7. From the figure, the feature distribution of positive and negative patches shows obvious differences, which also verifies the accuracy of the model in patch-level classification.

#### 4.4 Contribution

- (a) We only used slide-level labels (no pixels, patches, or ROI-level annotations) to train the model under weak supervision. Through weakly supervised training,



**Fig. 7** Visualizing the patch-level feature space. For each task (a, b, c), we use PCA to reduce the 512-dimensional feature representation to two dimensions (Left). For subtyping tasks (a, b, c), lesion area of chronic gastritis is clearly picked out by the positive cluster (Right)

relevant morphological features can be objectively identified from the micro-organization area without prior knowledge or subjective annotations. In the independent test, it is shown that the method in this paper can learn category-related morphological features, so it has the ability to recognize and diagnose.

- (b) Multi-label prediction capability. The  $n$  parallel attention branches constructed in Attention-MIL are well adapted to multi-label tasks. In the evaluation experiment, the advantages of this method in the multi-label prediction of gastritis pathological images are also proved, and the feasibility of weakly supervised learning in the multi-label task of pathological images is also proved.
- (c) Interpretability and interpretability as a clinical and research tool. We proved that our model is interpretable and can generate a heat map that can isolate the lesion area without using pixel-level annotations to identify the area containing the lesion. The patch-level classifier in the Attention-MIL can distinguish between positive and negative patches. Without pixel-level annotations, our method is able to identify the most relevant regions for classification determination.

## 5 Discussion

The identification of multiple pathological indicators not only involves an increase in the number of diagnostic tasks but also needs to deal with the relationship between multiple indicators. For example, “Activity” and “Atrophy” of chronic gastritis exist independently of each other, and the labels are in a parallel relationship. Therefore, in the design of the attention mechanism, we correspond the prediction of the attention network to the number of classifications  $n$ , thereby forming  $n$  parallel attention branches. Each branch will pay close attention to different tissue areas in WSI, and assign attention scores according to their importance to WSI level diagnosis. In this study, we set  $n = 3$  to achieve three classification tasks for chronic gastritis: “active”, “intestinal metaplasia” and “atrophy”. Each attention branch can learn which morphological features in each category contribute the most to the diagnosis of WSI level, to assign a higher attention score to it. Our research shows that in the evaluation of two independent test sets, the average accuracy and precision of the three diagnostic indicators are above 86%, the “activity” AUC is 0.951 and 0.929, the “intestinal metaplasia” AUC is 0.946 and 0.934, “atrophy” AUC is 0.926 and 0.898. These results prove that the  $n$  parallel attention branches constructed in the Attention-MIL model are well adapted to multi-label tasks and have unique advantages in the task of identifying multiple pathological indicators. Besides, we visualize the tissue area that Attention-MIL pays attention to and the distribution of two-dimensional features shows obvious differences, which fully shows that the Attention-MIL model can learn the corresponding morphological features of the category, so it has a high recognition ability.

## 6 Conclusions

With the development of AI technology in the field of pathological image recognition, it will bring innovative changes to the pathological diagnosis model. The Attention-MIL model proposed in this paper can assist in the identification of multiple pathological indicators at the same time, and further, improve the algorithm performance based on weakly-supervised learning. In the multi-index recognition of chronic gastritis, we evaluated the model on two independent test sets, thus verifying the correctness of the method in the multi-label task of gastritis pathological images. Visualizing suspicious diseased tissue regions has a certain explanatory power in clinical research, and it also provides new ideas for AI as a tool for assisting clinical diagnosis.

**Acknowledgements** This research work supported by the Shanghai Municipal Economic and informatics Commission’s project on artificial intelligence innovation and development 2019 (2019-RGZN-01017); Shanghai Pudong New Area Health System Important Weak Subject Fund (PWzbr2017-20).

**Data Availability** The data source of this article was provided by the partners Fudan University Cancer Hospital and Shanghai Pudong Hospital. this article mainly conducts multi-label prediction

research based on three pathological indicators of gastritis. the datasets used in this study are not publicly available due to specific institutional requirements governing privacy protection.

## References

1. Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A.: The 2019 WHO classification of tumours of the digestive system. *Histopathology* (2020)
2. Mukhopadhyay, S., Feldman, M.D., Abels, E., Ashfaq, R., Beltaifa, S., Cacciabeve, N.G., Cathro, H.P., Cheng, L., Cooper, K., Dickey, G.E., Gill, R.M., Heaton, R.P., Jr., Kerstens, R., Lindberg, G.M., Malhotra, R.K., Mandell, J.W., Manlucu, E.D., Mills, A.M., Mills, S.E., Moskaluk, C.A., Nelis, M., Patil, D.T., Przybycin, C.G., Reynolds, J.P., Rubin, B.P., Saboorian, M.H., Salicru, M., Samols, M.A., Sturgis, C.D., Turner, K.O., Wick, M.R., Yoon, J.Y., Zhao, P., Taylor, C.R.: Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *Am. J. Surg. Pathol.* **42**, 39–52 (2018)
3. Fuchs, T.J., Buhmann, J.M.: Computational pathology: challenges and promises for tissue analysis. *Comput. Med. Imaging Graph.* **35**, 515–30 (2011)
4. Liu, D., Bober, M., Kittler, J.: Visual semantic information pursuit: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1404–1422 (2019)
5. Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A.: Synthesize then compare: detecting failures and anomalies for semantic segmentation, pp. 145–161 (2020)
6. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666 (2019)
7. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019)
8. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018)
9. Golden, J.A.: Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *JAMA* **318**, 2184 (2017)
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017)
11. Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., Ding, X., Liu, J., Yu, C., Ku, C., Liu, C., Sun, Z., Xu, G., Wang, Y., Zhang, X., Wang, D., Wang, S., Xu, W., Davis, R.C., Shi, H.: Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat. Commun.* **11**, 4294 (2020)
12. Sharma, H., Zerbe, N., Klempert, I., Hellwich, O., Hufnagl, P.: Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput. Med. Imaging Graph.* **61**, 2–13 (2017)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
14. Huang, J., Li, Z., Li, N., Liu, S., Li, G.: AttPool: towards hierarchical feature representation in graph convolutional networks via attention mechanism. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6479–6488 (2019)
15. Caruana, R.: Multitask learning. *Mach. Learn.* **28**, 41–75 (1997)

16. Bragman, F., Tanno, R., Ourselin, S., Alexander, D., Cardoso, M.J.: Stochastic filter groups for multi-task CNNs: learning specialist and generalist convolution kernels (2019)
17. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **89**, 31–71 (1997)
18. Maron, O., LozanoPérez, T.: A framework for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **200**, 570–576 (1998)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
20. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: HoverNet: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019)
21. Zhou, Y., Graham, S., Koohbanani, N.A., Shaban, M., Heng, P., Rajpoot, N.: CGC-Net: cell graph convolutional network for grading of colorectal cancer histology images. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 388–398 (2019)
22. Cramme, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* **2**, 265–292 (2002)
23. Dubey, S.R., Chakraborty, S., Roy, S.K., Mukherjee, S., Singh, S.K., Chaudhuri, B.B.: Diff-Grad: an optimization method for convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 4500–4511 (2020)
24. Hoo, Z.H., Candlish, J., Teare, D.: What is an ROC curve? *Emerg. Med. J.* **34**, 357–359 (2017)

# A Survey of Object Tracking Methods Based on Deep Learning



Yang Yi , Zijian Meng , and Guixiong Tian

**Abstract** The object tracking problem is usually represented as estimating the state of an arbitrary target in a video only given its state in the initial frame. Many recent object tracking methods based on deep learning have achieved advanced performance. In general, most object tracking methods can be divided into two types. One family of trackers such as DiMP, ATOM learn a discriminative classifier, which adopt online updating strategy, to distinguish the target object from the background. In the process of tracking, these methods need to update the model parameters using reference frames and states. The other family of trackers represented by Siamese network use the reference state and the current frame as the input of the model without changing or updating the parameters. These trackers calculate the cross-correlation between the reference frame and the current frame to complete the object tracking task. Both methods usually divide object tracking task into two subtasks: classification and estimation. The target location is maintained by target classification, and the accurate target boundary box is obtained by target estimation. In this paper, we will investigate some state-of-the-art object tracking methods using deep learning in recent years. These methods will be introduced in three different emphases: classification, estimation, both.

**Keywords** Single object tracking · Deep learning · Computer vision

---

Y. Yi · Z. Meng · G. Tian

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China  
e-mail: [mengzj@mail2.sysu.edu.cn](mailto:mengzj@mail2.sysu.edu.cn)

Y. Yi

School of Information Science, Xinhua College of Sun Yat-sen University, Guangzhou, China  
Guangdong Province Key Laboratory of Big Data Analysis and Processing, Guangzhou, China

## 1 Introduction

Visual object tracking is a computer vision task, which is widely used in intelligent video surveillance, visual navigation, self-driving technology, human–computer interaction and other fields, aiming at analyzing video and tracking one or more categories of objects.

This paper studies the single object tracking problem, that is, in the whole tracking sequence, the tracking target is fixed and unique. The general single object tracking problem can be described as estimating the state of the target, including the position and size of the target, in a video given only the initial frame’s information. The state of target usually describes as a bounding box containing the position, height and width of the target.

In recent years, with the rise of deep learning (DL), many scholars have begun to use neural network to build their visual object tracking model. This paper reviews the single object tracking methods based on deep learning in recent years, and focuses on several advanced object tracking methods with different emphases. In Sect. 2, we will introduce the challenges, processes, common evaluation indicators and benchmark datasets of object tracking methods. In Sect. 3, we will detailedly discuss several visual object tracking methods based on deep learning. In Sect. 4, we will compare the methods to get their similarities, differences and advantages in various aspects. Finally, we will summarize the previous sections and draw conclusions, then put forward some views to the future work.

## 2 General Overview

In this section, we will describe the object tracking task. In Sect. 2.1, we will briefly introduce the difficulties and challenges of object tracking methods. In Sect. 2.2, we will introduce the process of target tracking methods. In Sects. 2.3 and 2.4, we will respectively introduce the common evaluation indicators and datasets of object tracking methods.

### 2.1 *Difficulties and Challenges*

The task of object tracking is to identify and track the specified target in the whole video sequence. Therefore, in the process of tracking, the target learned from the first frame usually needs to be constantly updated to adapt to the target changes caused by the illumination variation, occlusion, background clutter, scale variation, deformation, rotation, fast motion and out of view in the video sequence. However, how to effectively update the model is difficult, because in the process of object tracking, the target bounding box used for update is generated by the estimate module,

not the ground truth bounding box, which leads to the accumulation of tracking deviation. It will gradually make the estimated bounding box deviate from the target, resulting in the model drift.

In order to avoid the problem of model drift, a simple idea is to use the ground truth bounding box given in the initial frame to track the target all along. This is feasible in the ideal situation that targets do not change greatly, but it is obviously impossible. The video from natural scenes is likely to have huge deformation, using the most common situation as an example, the target moving away from or close to the camera will cause the target to become larger or smaller in the video, thus affecting the performance of object tracking. Moreover, how to ensure the speed of updates is also a challenging problem. Different from offline learning, model updating in object tracking is an online learning problem. In order to ensure the real time performance of object tracking, the time of model updating must be controlled in a short time. In addition, in order to achieve end-to-end training, the process of model updating must be embedded in the model, which is also a big challenge (Fig. 1).

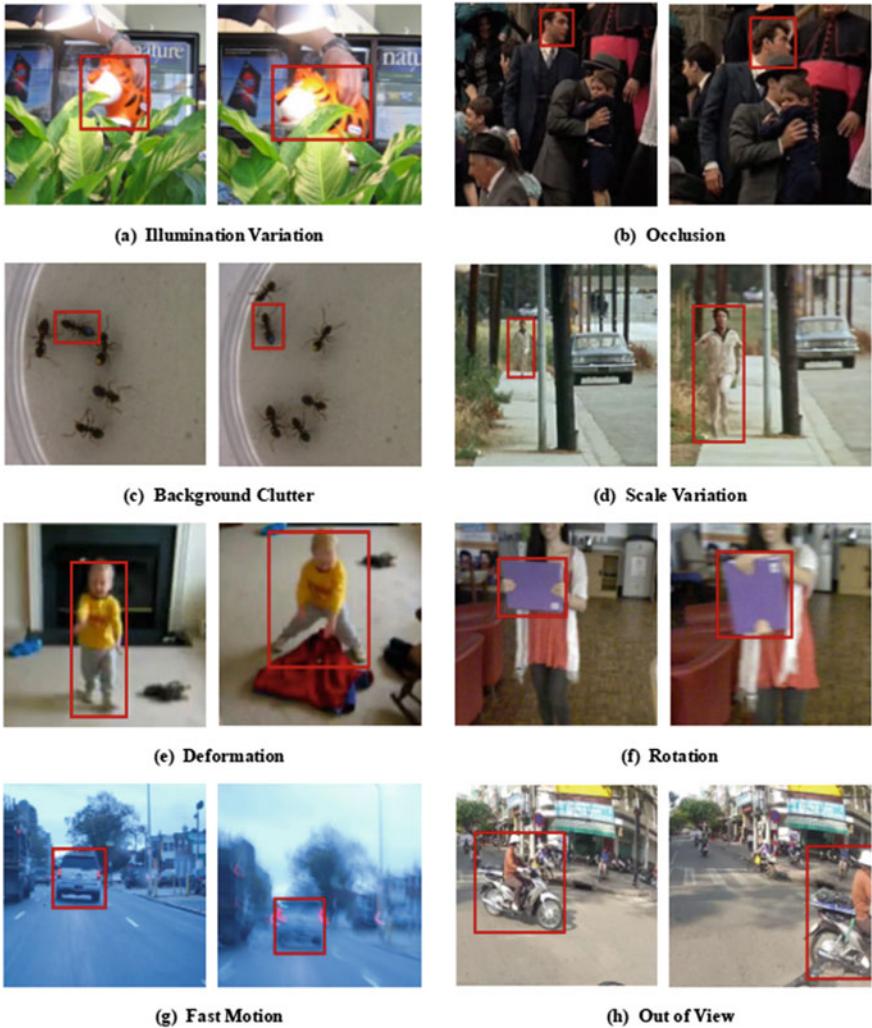
## 2.2 General Process

The task of visual object tracking can usually be regarded as the combination of two subtasks, classification and estimation. The classification task aims to determine the rough location of the target by classifying the foreground and background. The estimation task focuses on estimating an accurate target state that usually represented as a bounding box in the object tracking task.

In the task of visual object tracking, it is worth dividing the tracking process into object classification and bounding box estimation. An experiment by Bhat et al. [1] shows that robustness should be considered in the process of target classification, and high accuracy should be considered in the process of estimating target state. Target classification aims to find the location of the target, without good robustness, the target classifier may locate the background or similar distractors as the tracking target, which will cause the failure of the visual object tracking task. The purpose of boundary box estimation is to estimate the target state, which needs to cover the whole target without containing too much background. Therefore, the accuracy of estimation module is very important. Inaccurate boundary box estimation will increase the deviation when model updating and lead to more serious model drift.

Of course, some scholars [2–6] believe that completely splitting the two subtasks does harm to the performance of the tracker. Some of their researches [2, 4] have established more information exchange channels in the two subtasks to jointly solve the object tracking problem, and the other [3, 5, 6] see the two subtasks as parallel tasks.

In addition to the two subtasks, there is another important step in the object tracking method: feature extraction, which is usually depth feature extraction in the object tracking method based on deep learning. Feature extractor aims to extract the features of video frames through neural network, as the input of the classification



**Fig. 1** Difficulties and challenges in visual object tracking

### 2.3 Evaluating Indicators

There are five commonly used evaluating indicator in visual object tracking task:  $OP_T$  (overlap precision metric), AUC (area-under-the-curve), ACC (accuracy), Robustness, EAO (Expected Average Overlap).

- **OPT:** The ratio of the frames that the IoU score greater than the threshold  $T$  to all frames.
- **AUC:** The integral of the  $OP_T$  score that the threshold  $T$  from 0 to 1.
- **Accuracy:** The average overlap of the frames that track successfully.
- **Robustness:** The tracking failed ratio.
- **EAO:** Expected Average Overlap, that is, the integral of the average accuracy in a common length video.

Where accuracy, robustness and EAO are generally used as evaluation indicators for VOT series training sets,  $OP_T$ , AUC are commonly used in other datasets.

### 2.4 Datasets

There are several commonly used benchmark datasets in visual object tracking task: VOT (Visual Object Tracking dataset) [14], LaSOT [15], TrackingNet [16], GOT-10k [17], NFS (Need for Speed) [18], OTB-100 [19], UAV123 [20] and so on.

- **VOT:** This dataset consists of 60 challenging videos, Accuracy and Robustness are used to calculate EAO score to rank trackers.
- **LaSOT:** A large-scale dataset consists of 1400 videos with more than 3.5 M frames in total. The average video length of LaSOT is more than 2,500 frames, and each sequence comprises various challenges deriving from the wild where target objects may disappear and re-appear again in the view.
- **TrackingNet:** The first large-scale dataset and benchmark for object tracking in the wild. It provides more than 30 K videos with more than 14 million dense bounding box annotations, covering a wide selection of object classes in broad and diverse context.
- **GOT10k:** A massive dataset containing more than 10,000 videos, of which 180 videos constituted the test set for evaluation. There is no overlap of object classes between the training and test set, which increases the importance of generalizing unseen object classes. To ensure fair evaluation, trackers are prohibited from training with external datasets.
- **NFS:** The first higher frame rate video dataset and benchmark for visual object tracking. The dataset consists of 100 videos (380 K frames) captured with now commonly available higher frame rate (240 FPS) cameras from real world scenarios.

- **OTB-100:** dataset contains 100 videos with 11 challenges: Illumination Variation, Scale Variation, Occlusion, Deformation, Motion Blur, Fast Motion, In-Plane Rotation, Out-of-Plane Rotation, Out-of-View, Background Clutters, Low Resolution.
- **UAV123:** It consists of contains a total of 123 video sequences and more than 110 K frames, which are low-altitude aerial videos captured from unmanned aerial vehicles.

### 3 Methods

Visual object tracking methods usually decompose the target tracking task into two subtasks: target classification task and boundary box estimation task. In this section, we will introduce several target tracking methods with different emphases in the perspective of two subtasks.

#### 3.1 *Methods Focus on Target Classification*

Target classification, that is, by classifying the foreground and background to locate the target. In fact, there are many researches on object detector, so Wang et al. believe that some existing object detectors can be directly converted into high-performance trackers. It allows the tracker to retain its overall design. The main challenge here is to get the detector well initialized so that once a new tracking task is given, it can efficiently fine-tunes the tracker using target state without causing overfitting. MAML series network [21] are proposed based on this idea, they use meta learning method, puts forward a parameter initialization network. When offline training, it trains a well initialized parameters, so as to make the tracker have the ability to converge in a few steps in online update stage and then able to track the target. In addition, their network can be applied to any existing object detector to convert it to a target classifier, and their results can match the most advanced methods in object tracking task.

Another idea is to borrow some high-performance components from the existing advanced object detectors to build their own models. For example, Yang et al. also used the idea of meta learning and propose ROAM [8], which constructs an optimizer that can also converge quickly in online tracking stage based on meta learning. The DiMP network proposed by Bhat et al. [7] in 2019 has a model predictor, which predicts a filter by inputting the historical frames and their target bounding box. The filter convolutes with the current frame to generate the target response image and indicate the position of the target. In the offline training stage, the model predictor also uses the idea of meta learning to get an initial model which can converge quickly for all kinds of targets. And the steepest descent method is used to ensure the speed of training and online update, the most advanced result was obtained at that

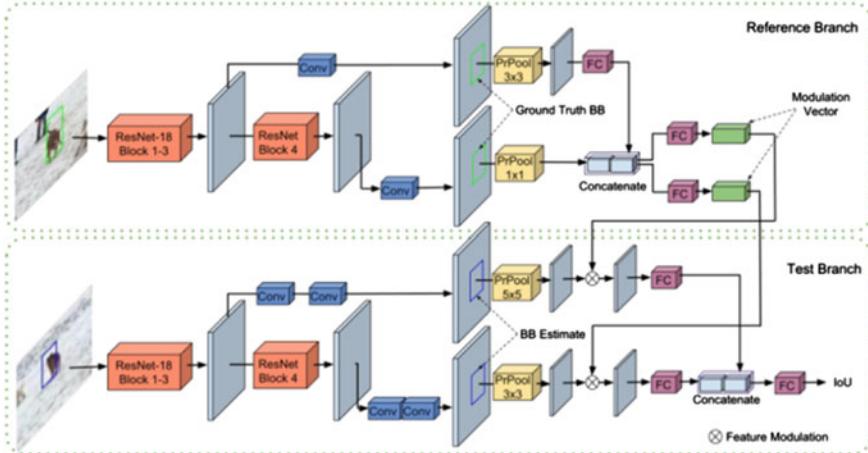
time. Daneljan et al., put forward PrDiMP [9] on the basis of DiMP, improved the target confidence response graph into the target probability response graph, giving a complete probability interpretation of the response graph, and achieved better results than DiMP.

### 3.2 *Methods Focus on Bounding Box Estimation*

In the aspect of bounding box estimation, the related research is relatively few. Many previous studies [11–13] use multi-scale anchor box searching strategy to estimate the bounding box. However, on the one hand, the regression speed and accuracy of this method are limited. On the other hand, it needs a lot of prior knowledge for anchor box setting. In addition, because it does not have enough high-level understanding of the target, its potential is limited. So Daneljan et al. proposed ATOM [10] to improve the accuracy and robustness of boundary box regression. Inspired by IoU net, ATOM takes rough bounding box as input and adjusts the bounding box parameters to maximize the predicted IoU score. In the offline training stage, the network parameters are updated to improve the accuracy of IoU score prediction. In the online tracking stage, the network parameters are fixed and the rough bounding box is used as a variable, then several iterations are carried out to maximize the IoU score. Finally, the predicted bounding box with the maximum IoU score is obtained. This method uses a simple target classifier with two convolution layers to complete the target classification task. In their ablation experiments, the target tracking effect is reduced by more than 10% after removing this simple target classifier. It can be seen that the target classification module has a great impact on the whole object tracking task. Compared with their previous optimal method, it has a relative improvement of more than 10% on each benchmark. It can be seen that a good boundary box estimator also has a huge improvement on the object tracking task (Fig. 2).

### 3.3 *Methods Focus on Both Subtasks*

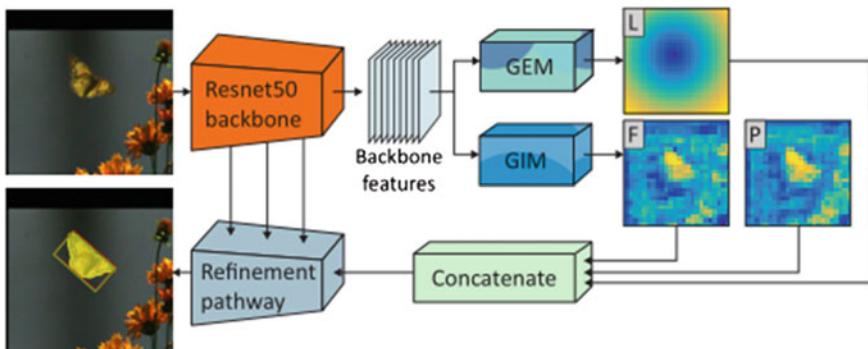
Some scholars believe that the separation of the two tasks will reduce the performance of the tracker, and miss the great opportunity to improve the tracking accuracy. For example, the D3S method proposed by Lukezic et al. [2] includes Gem (geometrically constrained Euclidean model) and Gim (geometrically invariant model) module. The former is used to generate robust target localization  $L$ , corresponding to the classification module in visual object tracking task, while the latter aims at calculating a posteriori exact target similarity graph  $P$  with calculating target similarity graph  $F$  (segmented real target and similar target) and background similarity graph  $B$ . Finally,  $L$ ,  $F$ ,  $P$  and the main features calculated above are input to the optimizer for joint solution, so as to obtain accurate pixel level target segmentation and generate



**Fig. 2** Full architecture of the target estimation network for ATOM [10]

a bounding box with angle, so that it can solve the target tracking task with high accuracy (Fig. 3).

Some other scholars [3–5] process the two subtasks at the same time. The SiamBAN method proposed by Chen et al. [3] regards the task of classification and estimation as a parallel task. They design a SiamBox module, which takes the each stages’ features of reference frame and the current frame as the input, and calculates the results of classification and regression in parallel. This module consists of a classification module *Cls*, which outputs a target confidence map with one channel, and a regression module *Reg*, which outputs a regression map with four channels (each channel represent a distance from this point to one of the sides of bounding box). Finally, the outputs of SiamBox module in each stage are adaptively fused to get an excellent tracking result.



**Fig. 3** The D3S segmentation architecture [2]

Xu et al. proposed a set of high-performance target tracker design guidelines for target tracking task, namely G1: decomposition of classification and state estimation; G2: non-ambiguous scoring; G3: prior knowledge-free; G4: estimation quality assessment. With the guidelines they designed a target tracker SiamFC++ without anchor boxes [4]. SiamFC++ also performs classification and estimation tasks in parallel, but different from SiamBAN, the two modules combine at the end to obtain a better result. By means of analysis and design of the classification problem, SiamFC++ achieve the state-of-the-art performance while maintaining the speed of 90 fps, although its structure is simple.

## 4 Experiment Results

This section will summarize the methods mentioned above in different datasets and scoring criteria. As shown in Tables 1 and 2.

We can see that in the VOT-2018 dataset, the D3S method has achieved the best results in each index, and the SiamFC++ series of methods have achieved good results with high FPS. It can be seen that the joint solution of classification and estimation tasks has its advantages.

While on other datasets, PrDiMP achieved the best results, which means that the target classifier with higher accuracy can provide a good improvement in the performance of the tracker. In addition, PrDiMP uses ATOM method as its estimation module so that it can achieve such a good result. From the point of view of improving performance, both tasks are very important for visual object tracking.

**Table 1** VOT-2018 benchmark dataset score

Method	EAO	Robustness	Accuracy	FPS
FCOS-MAML	0.392	0.220	0.635	42
Retina-MAML	0.452	0.159	0.604	40
DiMP-18	0.402	0.182	0.594	43
DiMP-50	0.440	0.153	0.597	43
PrDiMP-18	0.385	0.217	0.607	30
PrDiMP-50	0.442	0.165	0.618	30
SiamAttn	0.470	0.160	0.630	33
ATOM	0.401	0.204	0.590	30
D3S	<b>0.489</b>	<b>0.150</b>	<b>0.640</b>	25
SiamBAN	0.452	0.178	0.597	40
SiamFC++-AlexNet	0.400	0.183	0.556	<b>160</b>
SiamFC++-GoogLeNet	0.426	0.183	0.587	90

The best results are bold

**Table 2** AUC scores for several datasets

Method	Year	OTB-100	TrackingNet	LaSOT	NFS	UAV123
FCOS-MAML	2020	0.704	0.757	0.523	–	–
Retina-MAML	2020	<b>0.712</b>	0.698	0.480	–	–
ROAM	2020	0.681	0.620	0.390	–	–
ROAM++	2020	0.680	0.670	0.447	–	–
DiMP-18	2019	0.660	0.723	0.532	0.610	0.643
DiMP-50	2019	0.684	0.740	0.569	0.620	0.654
PrDiMP-18	2020	0.680	0.750	0.564	0.633	0.653
PrDiMP-50	2020	0.696	<b>0.758</b>	<b>0.598</b>	<b>0.635</b>	<b>0.680</b>
SiamAttn	2020	<b>0.712</b>	0.752	0.560	–	0.650
ATOM	2019	0.671	0.703	0.515	0.590	0.650
D3S	2020	–	0.728	–	–	–
SiamBAN	2020	0.696	–	0.514	0.594	0.631
SiamFC++-AlexNet	2020	0.656	0.712	0.501	–	–
SiamFC++-GoogLeNet	2020	0.683	0.754	0.544	–	–

The best results are bold

## 5 Conclusions

This paper introduces and compares the results of MAML, DiMP, PrDiMP, ATOM, D3S, SiamBAN and SiamFC++ which are the most advanced target tracking methods in recent two years. It can be concluded that:

- It is meaningful to divide the object tracking task into target classification task and boundary box estimation task. Even the methods that combine two tasks to solve the target tracking problem need to decompose the target tracking task into two subtasks.
- Target classifier is indispensable for target tracking task, even the simplest target classifier can provide a lot of performance improvement for target tracking task.
- A good boundary box estimation module can improve the performance of the algorithm greatly, whether in terms of speed, accuracy or robustness. Even with a simple target classifier, a good boundary box estimator can still make the performance of the whole algorithm reach the most advanced method.
- Combining two subtasks to solve the problem of object tracking can simplify the structure of the model and improve the speed and performance of target tracking.
- A set of feasible design guidelines is given, which can provide good design guidance for people, and a simple and easy-to-use target tracker is designed accordingly.

After a series of investigations, in my view, in the future the method of target tracking will tend to establish data path between two subtasks to solve the task of

object tracking jointly. In addition, the problem of long video tracking and the object tracking task that the target will be completely blocked remains to be solved.

**Acknowledgements** This work is partly supported by Guangzhou Science and Technology Project with No. 202002030273 and No. 201804010265, National Natural Science Foundation of China (NSFC No. 61672546), also by Key Discipline Project 2020XZD02, Xinhua college of Sun Yat-sen University.

## References

1. Bhat, G., Johnander, J., Danelljan, M., Khan, F.S., Felsberg, M.: Unveiling the power of deep tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–498 (2018)
2. Lukezic, A., Matas, J., Kristan, M.: D3S-A discriminative single shot segmentation tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7133–7142 (2020)
3. Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R.: Siamese box adaptive network for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6668–6677 (2020)
4. Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G.: SiamFC++: towards robust and accurate visual tracking with target estimation guidelines. Proc. AAAI Conf. Artif. Intell. **34**(07), 12549–12556 (2020)
5. Yu, Y., Xiong, Y., Huang, W., Scott, M.R.: Deformable siamese attention networks for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6728–6737 (2020)
6. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4282–4291 (2019)
7. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6182–6191 (2019)
8. Yang, T., Xu, P., Hu, R., Chai, H., Chan, A.B.: ROAM: recurrently optimizing tracking model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6718–6727 (2020)
9. Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7183–7192 (2020)
10. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: accurate tracking by overlap maximization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4660–4669 (2019)
11. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision, pp. 850–865. Springer, Cham (2016)
12. Sun, C., Wang, D., Lu, H., Yang, M.H.: Correlation tracking via joint discrimination and reliability learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 489–497 (2018)
13. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6638–6646 (2017)

14. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin Zajc, L., Sun, Y.: The sixth visual object tracking vot2018 challenge results. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0 (2018)
15. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Ling, H.: Lasot: a high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5374–5383 (2019)
16. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 300–317 (2018)
17. Huang, L., Zhao, X., Huang, K.: Got-10k: a large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* (2019)
18. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: a benchmark for higher frame rate object tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1125–1134 (2017)
19. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2411–2418 (2013)
20. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: European Conference on Computer Vision, pp. 445–461. Springer, Cham (2016)
21. Wang, G., Luo, C., Sun, X., Xiong, Z., Zeng, W.: Tracking by instance detection: A meta-learning approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6288–6297 (2020)

# Hyperspectral Image Classification Based on Stacked Extreme Learning Machine



Qiongying Fu , Jinchun Qin , and Li Li

**Abstract** Hyperspectral image classification has become a research focus in recent literature. However, extraction of deep features in a short time is still an open issue to increase classification performance. In this paper, combining the deep network structure model of deep learning and deep feature extraction with the fast network learning algorithm of extreme learning machine, a deep neural network model based on a stacked extreme learning machine is proposed. The network structure of the model contains multiple limits. First, the learning machine is constructed in an integrated cascading manner. The original features are initially learned using a multi-scale scanning strategy and used as model input. Second different attribution features learned by multiple extreme learning machines in each layer of the network are used as the input of the next layer of the network. And then separable attribution features are got for hyperspectral image classification. The experimental results demonstrate that the proposed method can obtain better classification performance.

**Keywords** Extreme learning machine (ELM) · Hyperspectral image classification · Deep feature extraction

## 1 Introduction

Hyperspectral remote sensing is an important technology of modern remote sensing. It can provide rich information in both spatial and spectral domains for the accurate identification and classification of ground objects. For hyperspectral image classification, extracting spectral features that can express and distinguish different feature categories is the key to increase the accuracy of image classification.

---

Q. Fu · L. Li

Department of Joint Operation, National Defense University, Beijing 100091, China

J. Qin

Department of Civil Engineering, Tsinghua University, Beijing 100084, China

e-mail: [qjc20@mails.tsinghua.edu.cn](mailto:qjc20@mails.tsinghua.edu.cn)

In recent years, scholars have conducted a lot of research on hyperspectral feature learning and classification based on neural networks. Ramón Moreno achieved classification by converting hyperspectral data into hyperspherical features and combining the extreme learning machine and Incremental extreme learning machine to obtain accurate thematic map of soybean crops [1]. Francisco Arguello and Dora B. Heras combined spectral features with extended morphological features or watershed features, and proposed a muti-kernel extreme learning machine that combines both spatial and spectral information [2, 3]. Koldo Basterretxea made effective use of the advantages of the ELM's fast computing and designed a real-time classification and processing plan for airborne hyperspectral images [4]. Although the extreme learning machine algorithm has a huge advantage in computing speed, its structure is still a single neural network.

Neural network with a deep network structure can better fit nonlinear functions and achieve complex function approximation. At the same time, it can learn the laws and characteristics of the target's deep potential from a large number of sample data. Professor Zhihua Zhou from Nanjing University proposed a new deep learning model integrated with multiple decision trees, called gcForest, with a forest composed of multiple decision trees as a single component [5].

The goal of this paper is to construct a network composed of multiple ELM using a structure like gcForest in order to get higher accuracy of classification with low time cost.

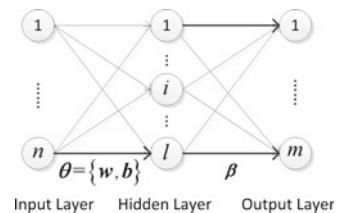
## 2 SELM Model Construction

### 2.1 ELM

ELM is an extreme learning algorithm for a single-layer feedforward neural network proposed by Huang [6]. The main idea is to randomly initialize the input weight and bias of the network model to find the optimal output matrix to minimize the deviation. The structure of ELM network is shown in Fig. 1.

As shown in the figure,  $n$ ,  $l$  and  $m$  are the number of nodes in the input layer, hidden layer and output layer respectively.  $\theta = \{w, b\}$  are the weight and bias between the

**Fig. 1** Structure of ELM



input layer and hidden layer, which are assigned through random initialization.  $\beta$  is the weight between the hidden layer and output layer.

Suppose there are  $N$  samples  $(X, T) = (x_i, t_i)$ ,  $i = 1, 2, \dots, N$ , where  $x_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T$ ,  $t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T$ . The output of the hidden layer corresponding to the input layer can be expressed as

$$\mathbf{H}(x_1, \dots, x_N; \theta_1, \dots, \theta_l) = \begin{bmatrix} g(x_1; \theta_1) & \cdots & g(x_1; \theta_l) \\ \vdots & \ddots & \vdots \\ g(x_N; \theta_1) & \cdots & g(x_N; \theta_l) \end{bmatrix}_{N \times l} \quad (1)$$

Then the output of ELM can be expressed as

$$f_l(x_i) = \sum_{j=1}^l \beta_j g(x_i; \theta_j) = \sum_{j=1}^l \beta_j g(a_j x_i + b_j) = \mathbf{H}\boldsymbol{\beta} \quad (2)$$

where the activation function is  $g(\cdot)$ ,  $g(\cdot)$  can be any bounded piecewise continuous function.

The process of training ELM is to solve the least squares solution of the linear system  $\min_{\boldsymbol{\beta}} \|\mathbf{H}\boldsymbol{\beta} - T\|$ . We can get

$$\boldsymbol{\beta}^* = \mathbf{H}^+ T. \quad (3)$$

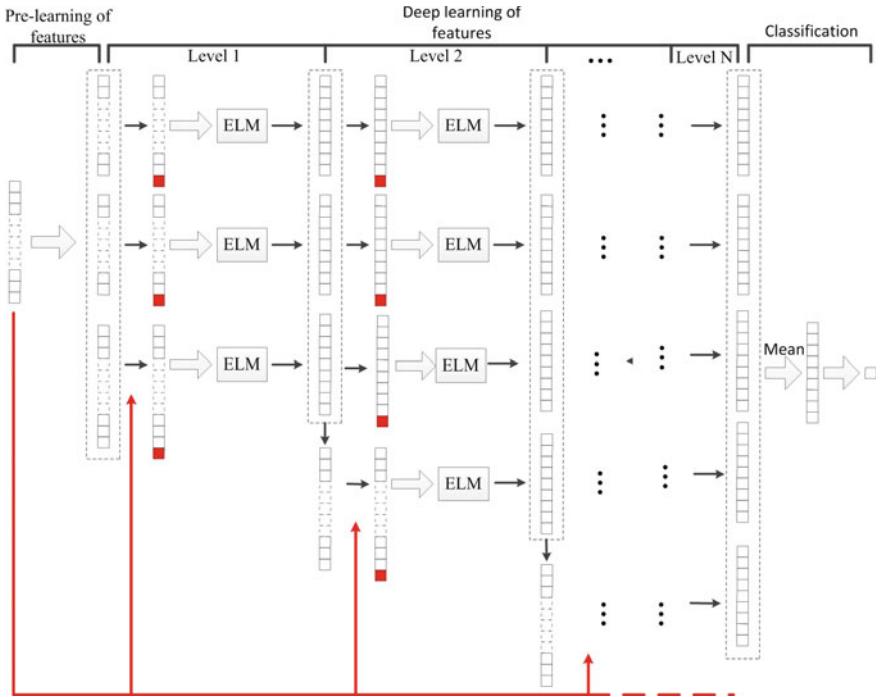
$\mathbf{H}^+$  is Generalized inverse of  $\mathbf{H}$ . It can be proved that  $\boldsymbol{\beta}^*$  has the smallest norm and it is unique.

## 2.2 SELM

The structure of SELM model adopts cascades of multiple extreme learning network. The model is mainly divided into three parts, feature pre-learning, feature deep learning and classification. Its structure is shown in Fig. 2.

**Pre-learning of features.** The input is the original feature vector and a certain strategy is used to divide the original feature into several sub-features in a certain way. Each sub-feature contains different feature information. The purpose is to construct different feature subsets to train the network from different perspectives.

**Deep feature learning.** This is the core part of the SELM model. The layers are connected in cascade mode, and the deep features are learned layer by layer. The output of pre-layer is the input of the next layer. Through learning of the input vector, deeper feature information can be get. In essence, each layer can be seen as



**Fig. 2** Structure of SELM

an integration of multiple different ELMs. The way of integration is to use different feature subsets as input.

Each ELM has a network which contains three layers. The number of nodes in the output layer is the number of categories of dataset to be classified. Take a dataset which has 103 spectral bands and 9 feature categories as an example. The number of nodes in the output layer of the ELM used to construct SELM is 9. For multi-categories classification tasks, the predictive output obtained by ELM is the weight of each neuron corresponding to each neuron category.

The category attribute weights learned by the input features are taken as the new features learned. For ease of description, such features are called attribute features. In order to strengthen the separability of features, two types of features are constructed as new inputs by using the attribution features of SELM input to each layer. One type is to cascade the attribution feature learned by each sub-feature with the original feature as a new feature; the other type is to cascade the attribution features learned from all sub-features first and then further cascade with the original feature as a new feature. For example, in the Level 1 stage in Fig. 2, three sub-features are generated, and the three ELMs will each generate a 9-dimensional attribution vector. Therefore, the input of the Level 2 network is three  $112 = 9 + 103$  dimensional features and a  $130 = 9 + 9 + 9 + 103$  dimensional features.

Assuming that the input vector is  $X$ , the sub-features obtained after the initial learning of the features are recorded as  $X_1^{L0}, \dots, X_n^{L0}$ , then the output of each layer of the network is

$$\begin{aligned} X_1^{L1} &= H(X_1^{L0})\beta \\ &\vdots \\ X_n^{L1} &= H(X_n^{L0})\beta \end{aligned} \quad (4)$$

The obtained features  $\tilde{X}_1^{L1} = [X_1^{L1}, X] \cdots \tilde{X}_n^{L1} = [X_n^{L1}, X]$  are cascaded with the original features  $X$ , and the inputs of the second-layer network are constructed as  $\tilde{X}_1^{L1} = [X_1^{L1}, X] \cdots \tilde{X}_n^{L1} = [X_n^{L1}, X]$  and  $\tilde{X}_{n+1}^{L1} = [X_1^{L1}, \dots, X_n^{L1}, X]$ , learned layer by layer.

**Classification.** Each ELM unit can get a class vector. In the first few layers of the network (except the last layer), the class vector obtained is not used as the classification result, but is used as a combination of the attributed feature and the original feature For the learning of deeper features. After the features of the last layer of the network are learned, the mean value of all attribution features is calculated, which is also a category vector, and the neuron label corresponding to the neuron with the highest output value is the category label predicted by the sample.

### 3 Classification Framework

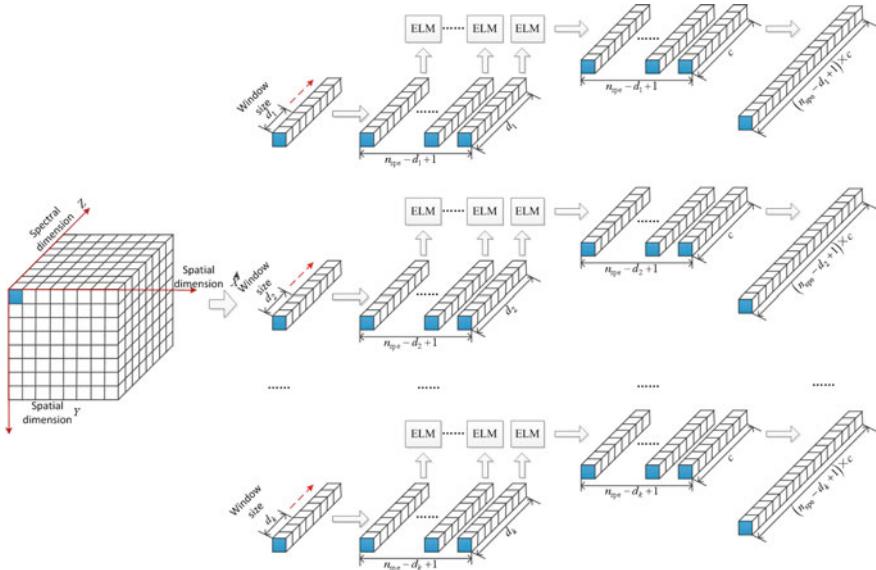
#### 3.1 Muti-scale Feature Scanning

In the typical deep learning model, when dealing with a common image problem, a single pixel in the image is often used as the processing unit, regardless of the relationship and function between adjacent pixels in the image. However, it can be seen from the existing research, the feature relationship between the original data can provide much more information for classification and recognition. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) algorithms can handle the relationship between data very well. For example, while processing two-dimensional images by CNN, convolution processing is performed on the original image, taking into account the image pixels and neighbouring pixels. RNN is very effective for sequence data. When dealing with speech recognition and other problems, the relationship between elements in sequence data and neighbouring elements is also considered.

For hyperspectral images, we only focus on spectral features here, which can be regarded as sequence data. There is correlation between bands. In order to use the spatial relationship between bands and adjacent bands, the multi-particle feature scanning method is used to learn sub-features in the feature learning part of SELM.

Use sliding windows of different sizes to scan the original spectral features. Assuming that the original spectral feature dimension is  $n_{spe}$ , define sliding windows, and the number of the windows is  $k$ , and each size is  $d_1, d_2, \dots, d_k$  respectively. Use each window to scan the original spectrum in turn to obtain the  $k$  sub-feature set of the group. The sub-feature set  $i (1 \leq i \leq k)$  of the first group contains sub-features  $n_{spe} - d_i + 1$ , and the size of each feature is  $d_i$ . Input the sub-features into the ELM model, and you can get  $k$  new set of sub-features. The new sub-feature set  $i (1 \leq i \leq k)$  in the first group also contains  $n_{spe} - d_i + 1$  sub-features, but the size of each feature is  $c$ ,  $c$  is the number of nodes in the ELM output layer, and is also the total number of categories of features. Concatenate the sub-features in each group to obtain eigenvectors  $k$ , and the size of each eigenvector is  $(n_{spe} - d_i + 1) \times c$ ,  $(1 \leq i \leq k)$ . This feature vector  $k$  is the overall sub-feature set learned in the initial feature learning stage, which serves as the input for subsequent deep feature learning. Figure 3 is the illustration of the multi-scale feature scanning method.

Also taking the dataset recommend in 2.2 as an example. If a sliding window with sizes of 20, 30, and 40 is used, in the initial feature learning stage, 3 sub-feature vectors are obtained through the multi-scale feature scanning method, the vector sizes are 756, 666, and 576 respectively.



**Fig. 3** The model of multi-scale scanning feature

### 3.2 Process of Classification Based on SELM

Based on the above principles, the specific implementation process of the SELM algorithm is summarized as follows.

**Input:** Hyperspectral image data, number of categories, window size.

**Output:** category label.

**Step 1:** Use windows of different sizes to scan to obtain group sub-features, input each group of sub-features to the extreme learning machines with output nodes in turn, and cascade the obtained output vectors.

**Step 2:** Input the cascaded vectors into the first-layer network as the initial learning features, and obtain the attribution features.

**Step3:** Cascade each attributed feature with the original feature separately to form the first type of new feature.

**Step 4:** After all attribution features are cascaded, they are cascaded with the original features separately to form the second type of new features.

**Step 5:** Take the two types of features together as the input of the next layer of the network to obtain new attribution features.

**Step 6:** Repeat Step 4 to Step 6 until the attribution feature of the last layer of network output is obtained.

**Step 7:** Calculate the average value of the attributed features of the output, and give the final category label of each pixel according to the weight.

## 4 Experiment Results and Analysis

This section designs two sets of experiments to verify and analyze the feature learning and classification of the stacked extreme learning machine proposed in this article. The experiments were conducted on two data sets. The first one is the Reno data set including 356 spectral channels in the 0.4–2.56  $\mu\text{m}$  region. The size of these data is  $313 \times 349$ .

The second data set is the Ibaraki data set. Due to the noise effects, the remaining 128 spectral channels were used in this paper. The size of the data set is  $650 \times 700$ .

During the experiment, 20% of each type of feature in each set of sample data is randomly selected as training samples, and the rest are test samples. Experiment 1 changed the number of network layers to build stacked extreme learning machines with different depths to evaluate the effectiveness of this algorithm for hyperspectral feature learning and classification; Experiment 2 used stacked autoencoders, extreme learning-based autoencoders, Convolutional neural network and stacked

extreme learning machine four deep network learning algorithms for classification and comparison analysis.

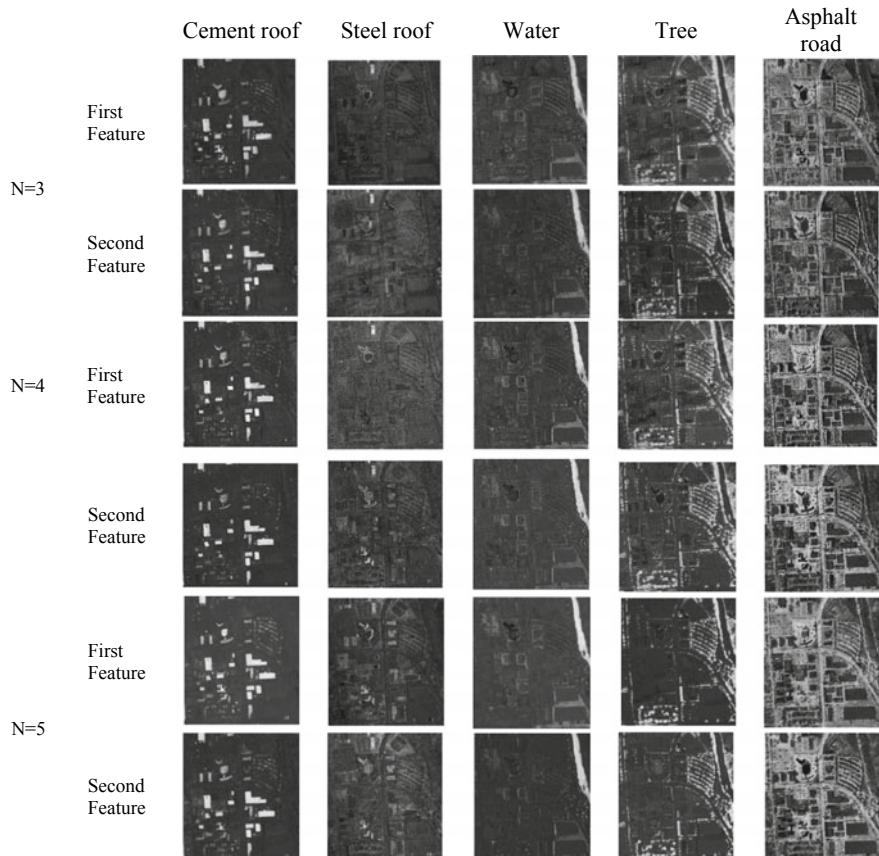
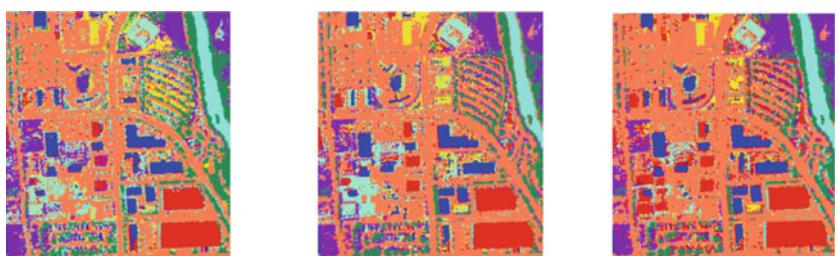
## 4.1 Deep Feature Learning

In the stacked extreme learning machine algorithm, the number of network layers determines the difference in the network structure. The expression performance of the hyperspectral image features learned by the neural network of different structures is different, which affects the ability to distinguish the types of features and Classification performance. This experiment changes the number of network layers and uses stacked extreme learning machines with different network depths for feature learning and classification. The number of layers here refers to the number of network layers in the deep feature learning part of the stacked extreme learning machine network structure. A single extreme learning machine is the basic unit as a layer, regardless of the three-layer network structure of the extreme learning machine network itself. During the experiment, five windows with sizes of 20, 30, 40, 50, and 60 were used for multi-scale feature scanning, and the number of network layers was set to 3, 4, and 5 in sequence. The value of each feature represents the weight of each feature's category. According to the learned features, the feature group map of the entire image can be obtained. Due to space limitations, only the front-end learning of different network structures can be obtained. The component diagrams corresponding to the two features are displayed, as shown in Fig. 4. In the figure, the brighter the grey value, the larger the proportion of a feature in the pixel, and the darker the smaller the proportion. Figures 5 and 6 show the classification results obtained by different networks. From Fig. 4, the attribution features learned by SELM can clearly show the attribution components between different features, and a better classification result is obtained, and the number of network layers increases, and the classification accuracy is improved.

## 4.2 Classification Performance

In order to verify the classification performance of the algorithm in this paper, a stacked autoencoder, an autoencoder based on extreme learning, and a convolutional neural network are used for classification comparison experiments. The parameter settings during the experiment are described as follows: SAE and ELM\_SAE each contain 5 hidden layers, the number of hidden layer nodes for Reno data is 1421, and the number of hidden layer nodes for Ibaraki data is 1041. The number of SELM network layers is 5. Tables 1 and 2 show the classification accuracy rates of the two sets of data under different algorithms.

From the experimental results in Tables 1 and 2, it can be seen that in terms of overall classification accuracy: the four deep learning algorithms can achieve higher

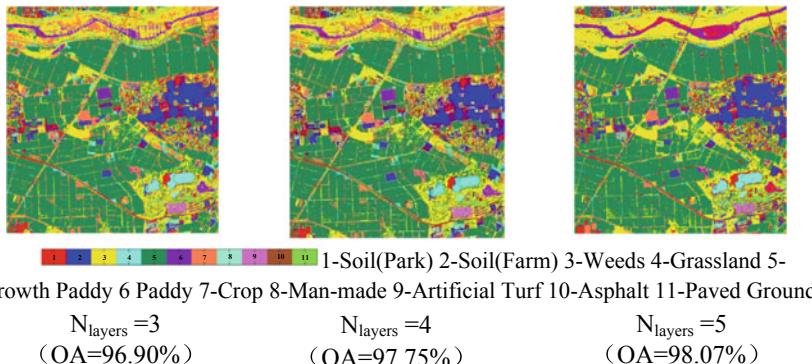
**Fig. 4** Features map for Reno data set

1-Tile roof 2-Cement roof 3-Steel roof 4-Water

5- Tree 6-Bare Soil 7-Asphalt road 8-Cement road

$N_{\text{layers}} = 3$  (OA=96.11%)  $N_{\text{layers}} = 4$  (OA=96.78%)  $N_{\text{layers}} = 5$  (OA=97.56%)

**Fig. 5** Classification maps for Reno data set



**Fig. 6** Classification maps for Ibaraki data set

**Table 1** Classification accuracy and training time using Reno data set (%)

Category	SAE	ELM_SAE	CNN	SELM
1 Tile roof	91.99	91.31	97.58	96.25
2 Cement roof	99.57	97.25	99.02	98.62
3 Steel roof	99.02	92.20	95.61	94.15
4 Water	99.76	99.76	99.71	99.51
5 Tree	98.95	98.07	97.19	98.59
6 Bare soil	99.90	100.00	100.00	99.58
7 Asphalt road	96.06	95.92	96.42	95.34
8 Cement road	97.09	97.62	98.15	98.41
Overall accuracy (%)	96.35	95.57	98.06	97.33
Training time (s)	4954.25	135.43	2301.45	218.86

classification accuracy, CNN has the highest classification accuracy, and the SELM algorithm proposed in this article is the second. Compared with SAE, ELM\_SAE uses ELM calculation to replace the iterative calculation in SAE's layer-by-layer initialization process during training time, which greatly shortens the training time. SELM also uses the calculation advantages of ELM, does not involve network parameter adjustment, and is equivalent to CNN. The accuracy of the classification is reduced while the calculation time is reduced, and the classification efficiency is improved.

**Table 2** Classification accuracy and training time using Ibaraki data set (%)

	Category	SAE	ELM_SAE	CNN	SELM
1	Soil (Park)	98.50	94.36	96.62	97.37
2	Soil (Farm)	88.74	85.96	93.70	91.31
3	Weeds	97.06	95.13	95.22	97.89
4	Grassland	99.36	98.99	99.50	99.41
5	Growth Paddy	99.82	99.54	99.96	99.98
6	Paddy	99.20	96.39	99.36	99.04
7	Crop	100.00	98.96	99.83	100.00
8	Man-made	100.00	97.06	100.00	97.06
9	Artificial turf	98.04	98.53	98.82	99.12
10	Asphalt	100.00	94.71	100.00	100.00
11	Paved ground	99.20	98.40	100.00	100.00
	Overall accuracy (%)	97.48	96.20	98.26	98.07
	Training time (s)	4178.64	147.65	2850.94	254.38

## 5 Conclusion

The experimental results with two hyperspectral data sets demonstrate that the deep features extracted by SELM can greatly improve the performance of hyperspectral image classification, with the sacrifice of reduced training time cost.

**Acknowledgements** The authors would like to thank America Lockheed Martin company and Headwall Photonics company for providing the Reno and Ibaraki data sets.

## References

1. Moreno, R., Corona, F., Lendasse, A., et al.: Extreme learning machines for soybean classification in remote sensing hyperspectral images. *Neurocomputing* **128**(5), 207–216 (2014)
2. Argüello, F., Heras, D.B.: ELM-based spectral-spatial classification of hyperspectral images using extended morphological profiles and composite feature mappings. *Int. J. Rem. Sens.* **36**(2), 645–664 (2015)
3. Heras, D.B., Argüello, F., Quesadabarriuso, P.: Exploring ELM-based spatial-spectral classification of hyperspectral images. *Int. J. Rem. Sens.* **35**(2), 401–423 (2014)
4. Basterretxea, K., Martinez-Corral, U., Finker, R., et al.: ELM-based hyperspectral imagery processor for onboard real-time classification. In: 2016 Conference on Design and Architectures for Signal and Image Processing. IEEE (2017)
5. Zhihua, Z., Ji, F.: Deep Forest: Towards An Alternative to Deep Neural Networks. (2017)
6. Guangbin, H., Qinyu, Z., Chee-Kheong, S.: Extreme learning machine: theory and applications. *Neurocomputing* **70**(1–3), 489–501 (2006)

# Nasopharyngeal Organ Segmentation Algorithm Based on Dilated Convolution Feature Pyramid



Xiaoying Pan<sup>1</sup> , Dong Dai<sup>1</sup> , Hongyu Wang<sup>1</sup> , Xingxing Liu<sup>1</sup> , and Weidong Bai<sup>1</sup>

**Abstract** In recent years, nasopharyngeal disease has been a common disease in clinical diagnosis whose incidence is increasing. As the most direct and effective means to observe the visceral mucosa of the cavity, Electronic laryngoscope, playing a role in diagnosis and minimally invasive diagnosis and treatment in the clinic, has become an important tool in otolaryngology head and neck surgeons. It is very important for clinical medicine to accurately segment the organs in the image. The following factors make organ segmentation more difficult, such as the complex structure and background of organs in nasopharynx and larynx; the unclear edge of organs and the little color differences between organs and background in laryngoscope image. In order to accurately segment organs and distinguish different instances of the same category, this paper proposes a nasopharyngeal organ segmentation model named Dilated Pyramid-Mask (DP-Mask). The model is based on the dilated convolution feature pyramid network. In my paper, Mask R-CNN is introduced into the organ instance segmentation of electronic laryngoscope image. What's more, in order to improve the segmentation accuracy, dilated convolution is designed in each layer of FPN to get the association of context feature information and get the segmentation of multiple organ instances in laryngoscope image. The experimental results show that the detection accuracy of DP-Mask model can reach 86.3%, Dice coefficient and mIOU can reach 0.81 and 0.88 respectively, which has high accuracy and robustness. Compared with popular U-Net and deep lab V3 algorithms, the proposed DP-Mask model improves mIOU by 5.4 and 4% respectively.

**Keywords** Instance segmentation · Mask R-CNN · Nasopharynx image · Dilation convolution

---

X. Pan · D. Dai · H. Wang · X. Liu · W. Bai

School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, China

e-mail: [panxiaoying@xupt.edu.cn](mailto:panxiaoying@xupt.edu.cn)

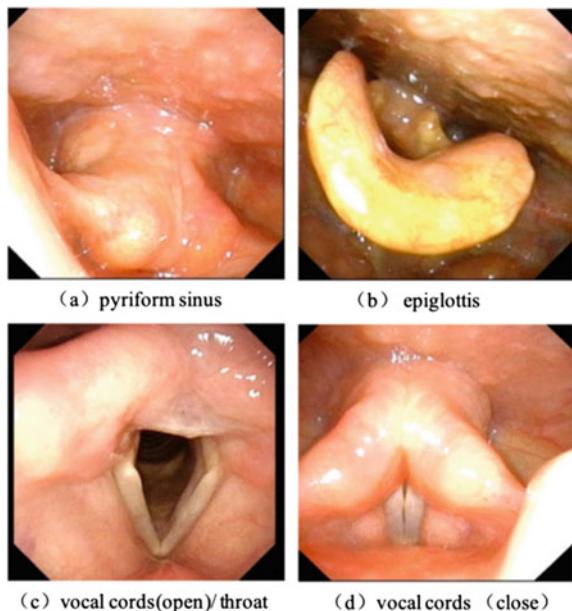
## 1 Introduction

In recent years, nasopharyngeal diseases are becoming more and more common, and their causes are diverse. Serious nasopharyngeal diseases will damage body health of patients anytime. Therefore, the medical research on nasopharyngeal diseases has gradually become a global medical problem. Electronic endoscope is an important tool for ENT doctors. It can display the image of electronic laryngoscope on the computer screen, in order to cut, mark and save the image, and assist doctors in laryngeal treatment [1].

Using manual interpretation to perform electronic laryngoscopy costs a lot of time and labor, and is prone to miss and misdiagnosis, so it is necessary to rely on computer-assisted doctors for diagnosis. The various organs of the nasopharyngology can be segmented by image segmentation technology, which can help doctors accurately judge whether there are lesions in this part. The objects studied in this paper mainly include epiglottis, vocal cord, piriform fossa, larynx and other organs, as shown in Fig. 1. These organs have a very high incidence in the clinical diagnosis of otolaryngology department at present. Accurate segmentation of these organs can determine the specific location of lesions, and help doctors to analyze whether these organs are deformed due to hyperplasia, cyst and other lesions. Therefore, it is very important to accurately segment relevant organ parts from electronic laryngoscope images.

With the breakthrough of deep learning technology in the field of computer vision, image segmentation has shown the intelligent effect closest to human expectation [2].

**Fig. 1** Organs of nasopharynx

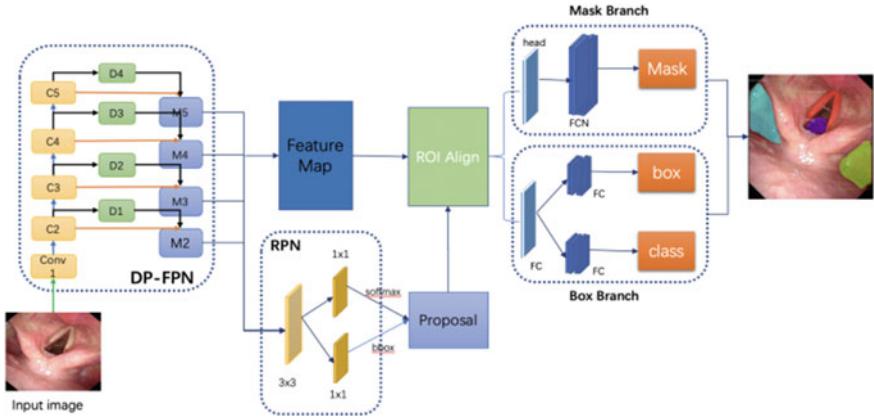


At present, the application of this technology in endoscopic images is mainly used in gastroscopy, colonoscopy and capsule endoscopy for early cancer screening and bleeding detection [3]. Tomoyuki et al. [4] used Mask R-CNN to detect endoscopic image data and segment gastric cancer region. In the segmentation effect evaluation, the average Dice index reached 71%. Some researchers applied the DeepLab network to the segmentation task of medical images. For example, Wang et al. [5] added an optimal dense prediction unit to the DeepLab 3+ network, which significantly improved the detection accuracy of intestinal polyps under colonoscopy. In addition, DeepLab v3+ network has also been applied to the segmentation of early esophageal cancer in endoscopic images [6]. Zheng et al. [7] improved the U-Net network and designed a two-channel separation convolution model, which effectively improved the learning ability of the network model on multi-scale features. In the segmentation task of colorectal cancer tumor, compared with U-Net network, the Dice value increased by 6.42%. This study effectively demonstrated that multi-scale feature fusion played an important role in improving the accuracy of lesion segmentation. In addition, in the segmentation of gastric cancer pathological images, this conclusion was further proved by literature [8]. Wittenberg et al. [9] applied Mask R-CNN to colonoscopy image data to detect colon polyps, and the results were 0.86, 0.80 and 0.74 on three open data sets. Tomoyuki et al. [4] used Mask R-CNN to detect endoscopic image data and segment gastric cancer region. In the segmentation effect evaluation, the average Dice index reached 71%.

The nasopharyngology image data studied in this paper mainly has the following characteristics:

- (1) The cavity phase structure of the nasolaryngology is complex, and the internal tissues are of different shapes, most of which are convex or concave.
- (2) The color features of nasolaryngology images are not obvious, the contrast is low, and the pixel difference between organs and background is small;
- (3) The boundaries of organs in the nasolaryngology are not clear and difficult to distinguish;
- (4) Due to the different operation techniques of different examiners, the imaging effects of different endoscopes are different, and there are great differences between different data.

In view of the above problems, traditional segmentation methods could not achieve satisfactory segmentation results. In this paper, an improved framework based on Mask R-CNN, DP-Mask (Dilated Pyramid Mask), was proposed to complete the example segmentation of nasopharyngology images. In this method, Feature Pyramid Networks (FPN) [10] was used to extract multi-scale Feature information in the Feature extraction stage, and dilated convolution was added to the Feature Pyramid to obtain more global information. This method not only improves the fusion ability among multi-scale feature levels, but also makes up for the serious and irreversible loss of image information in the process of up-sampling, which is very important for improving the accuracy of image segmentation. At the same time, in the training process, random rotation and mirroring are used to expand the data and improve the robustness of the model.



**Fig. 2** Basic frame diagram of DP-Mask

## 2 DP-Mask Segmentation Model

### 2.1 Model Framework

Based on the data characteristics and difficulties in segmentation of nasolaryngology images, a two-stage example segmentation model, DP-Mask, is proposed in this paper. In the first stage, the images are scanned and the proposed Proposals are generated. And in the second stage, the Proposals are classified and the boundary frames and masks are generated. The model mainly includes Feature extraction module DP-FPN (Dilated Feature Pyramid Networks), region generation module, region feature aggregation module, FCN segmentation module [11], etc. In addition, dilated convolution is introduced into the Feature pyramid to obtain more global information. It improves the ability of fusion among multi-scale feature levels. At the same time, in the process of training, the samples are enhanced by random flipping, color jittering, noise interference and other enhancement techniques to realize data dilated and improve the robustness of the model [12]. The algorithm frame diagram is as follows (Fig. 2).

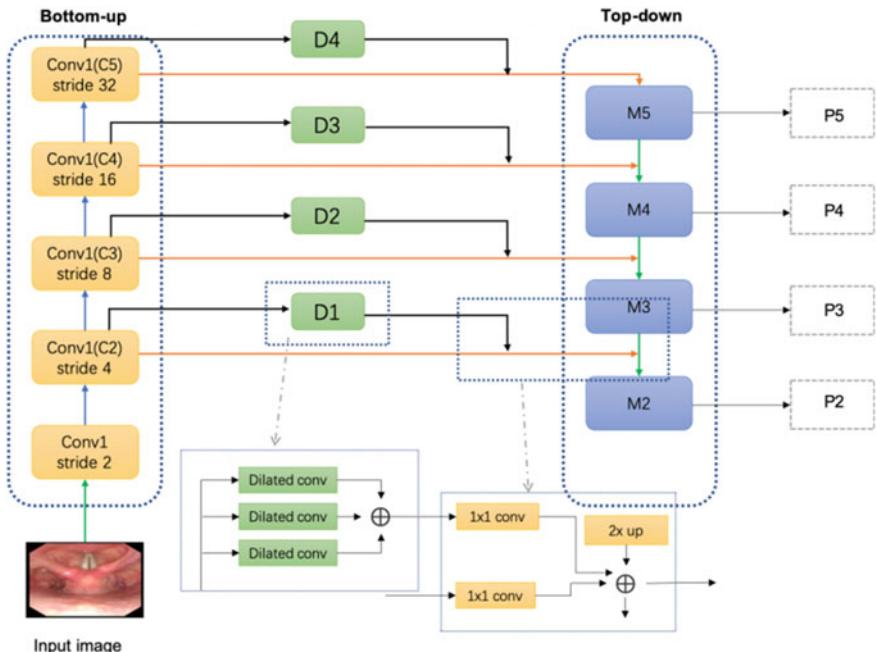
### 2.2 Multi-scale Feature Extraction Module

Multi-scale feature extraction [13] is becoming more and more important in the target detection task. It can improve the accuracy in the field of image segmentation. However, in the process of first decreasing the size and then enlarging the size, the image loss is serious, which affects the prediction accuracy of the Mask.

The proposed feature extraction module DP-FPN (Dilated Pyramid FPN) outputs features at four scales: P5, P4, P3 and P2. Firstly, the image is continuously subsampled to obtain features at five scales C1, C2, C3, C4 and C5, and then continuously up-sampled from C5 to obtain features M5, M4, M3 and M2. The features at each scale obtained in the process of up-sampling are fused with those at the corresponding scale in the process of down-sampling, that is, horizontal connection. Three dilated convolution operations are used in each horizontal connection. The convolution kernel is  $3 \times 3$ , and the dilated rates are 2, 4, and 8, respectively. The features obtained by dilated convolution are fused with the transverse connection features to obtain a new horizontal feature, and the structure is shown in Fig. 3. By introducing the dilated convolution operation, the receptive region of the convolution kernel can be increased; the information loss in the process of using the feature pyramid can be reduced, and the fusion ability of deep features and shallow features can be improved, so as to improve the accuracy of organ segmentation.

Expanding convolution is to add some weight of 0 to the convolution kernel on the basis of the standard convolution operation, so that each convolution output can contain a large range of information. The expression is as follows:

$$RF_i = RF_{i-1} + (k - 1) \times s \quad (1)$$



**Fig. 3** Structure diagram of DP-FPN

$RF_{i-1}$  is the receptive field of the upper layer, K is the size of the convolution kernel, and S is the step size.

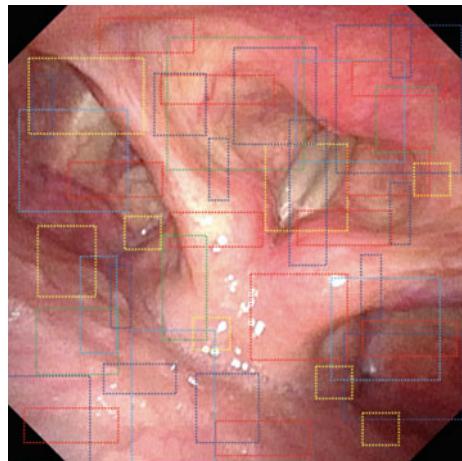
### 2.3 Region Generation Module

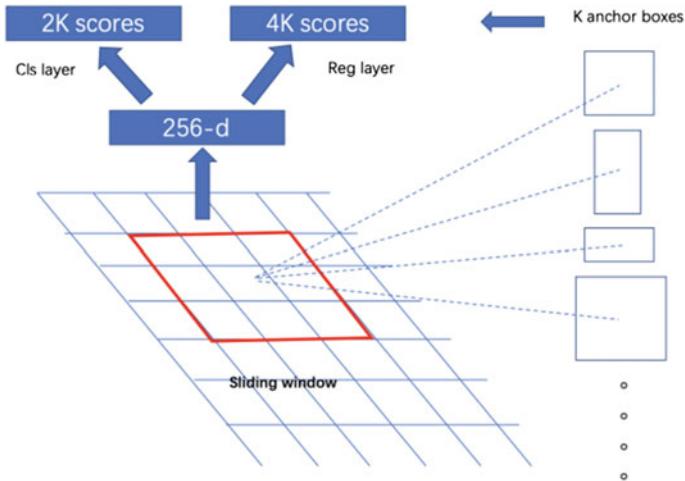
The region generation module is a lightweight neural network, which scans the feature map with the sliding window and looks for the existing target region. After scanning, rectangular regions of different sizes and length-to-width ratios will be generated on the feature and become anchors, usually nearly 20,000. These anchors can overlap and contain each other. And you want to cover the entire image as much as possible. As shown in Fig. 4.

The input of this module is the multi-scale feature map generated by DP-FPN, and the output is the category and finely tuned border of Anchor respectively. Here, only the foreground or background of Anchor can be distinguished. If the foreground is classified, it is considered that there is probably a target in the Anchor box. As for the position of the anchor box, sometimes it is not perfect at the center of the target, so it needs to be corrected, which can be realized by translation or scaling. When the anchor is close to the real box, it can be fine-tuned by using advanced regression. The structure of region generation module is shown in Fig. 5.

Using this area to generate modules, we can select the best anchor that contains target and fine-tune its position and size. If there are multiple anchors overlap each other, non-maximum Suppression (NMS) [14] algorithm can be used to eliminate redundant anchors and only the one with the highest prospect score was retained. Then we got the final regional proposal and passed it on to the next stage.

**Fig. 4** Sample diagram of Anchor coverage





**Fig. 5** Area generated module structure diagram

## 2.4 Regional Feature Aggregation Module

After going through the LAN, we get a series of RFPs. These Bboxes are usually of different sizes, which is about 2 K, and we need to extract equal-sized, fixed features from these different Bboxes. In this method, bilinear interpolation method is introduced to obtain the image values on the float point coordinate pixels based on ROI Pooling. The whole feature aggregation process is a continuous operation. The algorithm flow is as follows:

- (1) Each candidate region is traversed so that the non-integer boundary is not quantized;
- (2) Divide the candidate region into  $k * k$  subregions, and the boundary of the subregions is not quantized;
- (3) Divide each subregion into four squares again, and use bilinear interpolation method to calculate the center point of the small square as the value of the square, so that each subregion has four such values;
- (4) The maximum value of the four values obtained in the previous step is taken as the value of each subregion, so  $k * k$  values are obtained, and the result is output as a feature graph.

### 3 Experiment Part

#### 3.1 Data Origin

In this study, the image data of electronic laryngoscope from the Department of Otolaryngology of a “AAA” hospital in Xi’an were used. Through the video acquisition card directly from the electronic endoscope workstation pull video streaming as raw data. The collecting cycle was one month, and a total of 350 video data were collected from different patients. The video ranging from two to five minutes long is used as the data set of this experiment through the data further and processing.

#### 3.2 Data Preprocessing

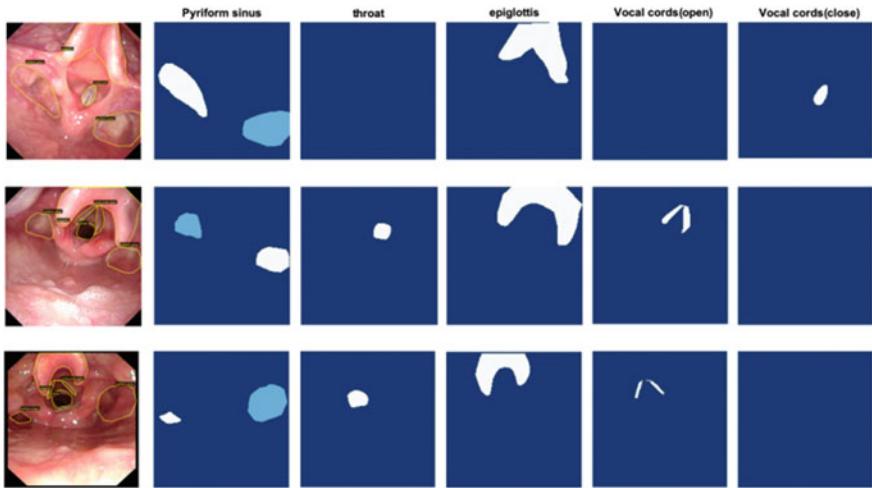
From the 350 video data collected, 250 samples were randomly selected as the training set and 50 samples were randomly selected as the verification set, and the other 50 samples were selected as the test set. First, the video data was dismantled to filter out invalid fragments such as external images and fuzzy images, and the image resolution was  $1060 * 1075$  after the useless border of the original image was clipped. By extracting key frames as samples, we obtained a total of 8000 images of training set, 1000 images of verification set and 1000 images of test set. The training samples were tagged using the VIA image tagging tool. The whole tagging work was completed with the assistance of the physician and was reviewed and confirmed by another senior otolaryngologist. The annotation method is to use a series of points to mark the outline of the target image. In Fig. 6, the label based on the original image is on the far left, and the mask of each organ part is on the right.

In order to overcome the overfitting phenomenon of neural network and improve the robustness of the model, this paper adopts random flipping, color dithering, noise interference and other enhancement techniques. One enhancement method was adopted randomly for each training sample, so that the size of training set and verification set were both doubled, with a total of 16,000 training sets and 2000 verification sets.

#### 3.3 Training and Implementation

The model of this experiment is built by Keras and TensorFlow [15] framework. The experimental platform is Linux cloud server, and the hardware is: Intel Xeon E5-2620 CUP, RAM 128G, and Tesla K80 GPU\*2.

During training, all images were scaled to  $416 * 416$ , and 8 images were sent into the model for training. 300 steps were trained for each epoch, 20 steps were verified, and a total of 100 epochs were trained. The initial learning rate was set at 0.001, the



**Fig. 6** Data set annotation format and mask

attenuation weight was 0.0001, and the non-maximum inhibition IOU threshold was 0.5.

In order to verify the performance of the segmentation method in this paper, the measurement standards commonly used in medical image segmentation are adopted in the experiment: Average Precision (AP), Dice coefficient, IOU, etc.

### 3.4 Experimental Results and Analysis

In order to verify the effect of different feature extraction networks on the method in this paper, this experiment trained the model based on Resnet-50-FPN and Resnet-101-FPN respectively, and the results were shown as follows (Table 1).

Experimental results show that the method in this paper has a better effect in Resnet-50-FPN, because the experimental data is limited in this paper; the color in the image is single; the noise interference is small; the semantic information of the target is obvious; the semantic information of the target is less dependent on the high-level semantic information. Therefore, We don't need a larger Resnet-101 with better

**Table 1** Comparison of AP values in different backbone networks

Backbone	Ap50	Ap60	Ap70	Ap80	Ap90
Res-50-FPN	<b>0.863</b>	<b>0.805</b>	<b>0.696</b>	<b>0.464</b>	<b>0.161</b>
Res-101-FPN	0.835	0.756	0.660	0.448	0.137

Boldface means that the data is best compared to the other comparison data in the table

**Table 2** Comparison of IOU and dice coefficients of each organ

Organ	IOU-res50	IOU-res101	Dice-res50	Dice-res101
Epiglottis	<b>0.87</b>	<b>0.78</b>	0.86	0.82
Pyriform sinus	0.69	0.66	0.71	0.75
Throat	0.77	0.74	0.84	<b>0.83</b>
Vocal cords (open)	0.71	0.64	0.79	0.76
Vocal cords (close)	0.82	0.77	<b>0.88</b>	<b>0.83</b>
mIoU	0.772	0.718	–	–
mDice	–	–	0.816	0.797

Boldface means that the data is best compared to the other comparison data in the table

high-level feature extraction capabilities, for Resnet-50 can achieve better results. This also illustrates the point that the deeper the network may not be a good method.

In this experiment, the segmentation effect of a single organ was also statistically analyzed. IOU and Dice coefficients were used to evaluate the segmentation effect. The experiments were conducted on RES-50 and RES-101 feature extraction networks, and the experimental results were shown in Table 2.

The results showed that the segmentation model based on Resnet-50 had a better segmentation effect on vocal cords (closed), larynx, epiglottis and other organs. The IOU of epiglottis reached 0.87, and the Dice coefficient of vocal cords (closed) was the highest, reaching 0.88. The second is epiglottis and larynx, and the Dice coefficient reaches 0.86 and 0.84, respectively. By comparison, the segmentation effect of piriform pit is the worst. Through the observation and analysis of the image, there is an obvious color difference between the vocal cords and the background, which is easy to distinguish. Although the area of larynx is smaller than that of other organs, its shape is regular and uniformly round. The area of epiglottis is relatively large, taking a large proportion in the whole image, and it is elliptical. However, the poorly segmented part of piriform fossa presents different shapes from different angles because there is no clear organ edge and no regular shape, which brings great difficulty to the segmentation.

The following figure (a) shows the IOU confusion matrix of mask prediction based on the Resnet-50 model, and (b) shows the IOU confusion matrix of BBOX based on the Resnet-50 model. In the region generation module, the accuracy of the candidate box has a great influence on the segmentation results, which is also proved in the experimental results. In the figure, the prediction accuracy of the Bbox of the piriform fossa and larynx is relatively low, and the prediction accuracy of the corresponding Mask is relatively low, while the prediction accuracy of the Bbox of the epiglottis and vocal cords (closed) is relatively high. The corresponding Mask accuracy is also high (Fig. 7).

The following figure shows the segmentation effect comparison between this experimental model and other mainstream image segmentation algorithms. From the figure, it can be observed that the segmentation effect of this experimental method is significantly better than other methods. U-Net segmentation model cannot

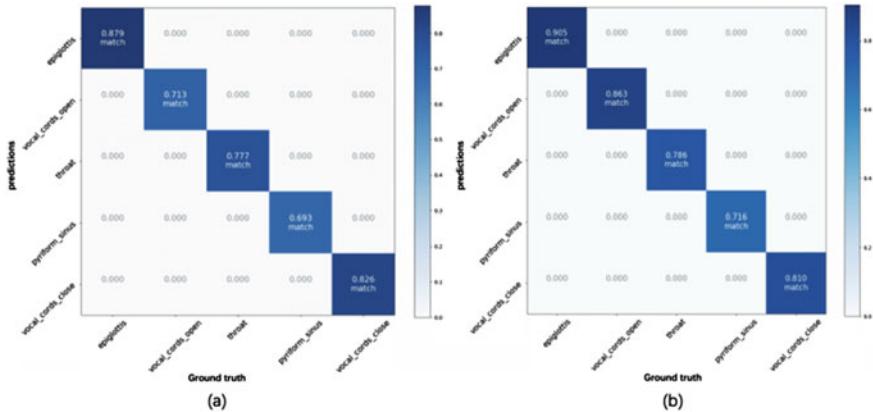


Fig. 7 Data set annotation format and mask

segment organs completely, with uneven segmentation edges and internal defects. The segmentation effect of Yolact model is better, but there will be misdetection (Fig. 8).

In order to verify the effect of the method in this paper, the experiment compares the method in this paper with Yolact, Mask R-CNN and other example segmentation models. Based on the COCO data set format, In the experiment we use the average precision AP of different thresholds to evaluate the effect. The results are shown in Table 3.

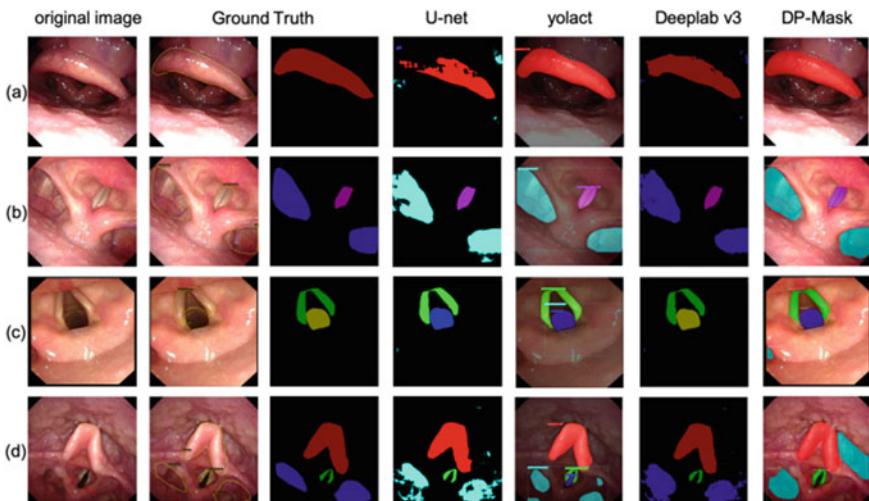


Fig. 8 Visualize the results of each model

**Table 3** Comparison of AP values for instance segmentation

Model	Ap50	Ap60	Ap70	Ap80	Ap90
Yolact	0.871	0.763	0.613	0.311	<b>0.277</b>
Mask R-CNN	0.701	0.632	0.491	0.280	0.032
DP-Mask	<b>0.863</b>	<b>0.805</b>	<b>0.696</b>	<b>0.464</b>	0.161

Boldface means that the data is best compared to the other comparison data in the table

**Table 4** Semantic segmentation results IOU comparison of different organs

Organ	Mask R-CNN	U-Net	DeepLab v3	DP-mask
Epiglottis	0.68	0.75	0.74	<b>0.87</b>
Pyriform sinus	0.57	0.49	0.56	<b>0.69</b>
Throat	0.69	0.70	0.69	<b>0.77</b>
Vocal cords (open)	0.55	0.64	0.65	<b>0.71</b>
Vocal cords (close)	0.64	0.81	0.80	<b>0.82</b>
mIoU	0.64	0.68	0.69	<b>0.77</b>

Boldface means that the data is best compared to the other comparison data in the table

The results show that the effect of the proposed method is significantly improved compared with before. From AP50 to AP80, the proposed method is better than the other two example segmentation models, in which AP50 and AP90 are 0.16 and 0.13 higher than that of the Mask R-CNN.

Table 4 shows the results of the proposed method compared with the original Mask R-CNN and other major semantic segmentation methods. Based on the Pascal VOC data set format, the mIOU of the proposed method is 0.772, which is 0.14 higher than that of the Mask R-CNN, and the epiglottis and vocal cord parts have the most obvious improvement. Compared with other semantic segmentation models such as U-Net and DeepLab V3, the results are 8% and 7% higher, respectively.

We introduced dilated convolution into FPN, and use different dilated factors at different scales which not only improves the receptive field of the convolution kernel, but also enhances the correlation between high-level features and low-level features, making up for the information lost in the process of up-sampling, thus improved the effect of image segmentation.

## 4 Conclusion

In the screening of otolaryngology, accurate segmentation of nasopharynx organ parts is of great help to the diagnosis. In order to accurately segment the parts of organs, clearly outline the contour of organs, and distinguish different instances of organs,

this paper proposed an improved instance segmentation model based on Mask R-CNN and named DP-Mask. In the model, the dilated convolution is used to amplify the receptive field of each horizontal connection in the FPN, and then it is added to the corresponding layer features in the process of upper adoption, which greatly avoids the information loss of the traditional FPN in the upper sampling node, and introduces the context information for each layer of the feature pyramid. The model achieves a good segmentation effect on the electronic laryngoscope image, with the mIOU reaching 0.772 and the Dice coefficient reaching 0.816 in the test set.

**Acknowledgements** First of all, I would like to express my deepest gratitude to my supervisor, Professor Pan Xiaoying. Thanks to her constant encouragement and guidance, she guided me through every stage of the paper.

I would also like to thank Ms. Wang Hongyu for her help in writing the paper. Under her patient guidance, I finished the writing of the paper smoothly.

This work was supported by the National Natural Science Foundation of China (Program No. 62001380).

## References

1. Peng, X., Youlin, Q., Yu, J.: Application of artificial intelligence in diagnosis of medical endoscope. *Zhong Hua Zhong Liu Za Zhi Chin. J. Oncol.* **40**(12) (2018)
2. Xiangbin, L., Liping, S., Shuai, L., et al.: A review of deep-learning-based medical image segmentation methods. *Sustainability* **13**(3), 1224 (2021)
3. Atsuo, Y., Ryota, N., Keita, O., Tomonori, A., Kazuhiko, K.: Automatic detection of colorectal neoplasias in wireless colon capsule endoscopic images using a deep convolutional neural network. *Endoscopy* (2020)
4. Tomoyuki, S., Atsushi, T., Hyuga, Y., et al.: Automated detection and segmentation of early gastric cancer from endoscopic images using mask R-CNN. *Appl. Sci.* **10**(11), 3842 (2020)
5. Yagang, W., Yiyuan, X., Xiaoying, P.: Method for intestinal polyp segmentation by improving DeepLabv3+ network. *J. Front. Comput. Sci. Technol.* **14**(07), 1243–1250 (2020)
6. Dingyun, L., Hongxiu, J., Nini, R., et al.: Computer aided annotation of early esophageal cancer in gastroscopic images based on DeepLabv3+ network. In: Proceedings of the 2019 4th International Conference on Biomedical Signal and Image Processing (ICBIP 2019), pp. 56–61(2019)
7. Suichang, Z., Xue, L., Weifeng, Z., et al.: MDCC-Net: multiscale double-channel convolution U-Net framework for colorectal tumor segmentation. *Comput. Biol. Med.* **130**, 104183 (2020)
8. Feng, G., Canghong, S., Xiaojie, L., Xi, W., Jiliu, Z., Jiancheng, L.: Image segmentation of nasopharyngeal carcinoma using 3D CNN with long-range skip connection and multi-scale feature pyramid. *Soft Comput.* (2020)
9. Wittenberg, T., Zobel, P., Rathke, M., et al.: Computer aided detection of polyps in whitelight-colonoscopy images using deep neural networks. *Curr. Dir. Biomed. Eng.* **5**(1), 231–234 (2019)
10. Tseng-Yi, L., Piotr, D., Ross, G., et al.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944 (2017)
11. Evan, S., Jonathan, L., Trevor, D.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2015)
12. Maro, M.K., Su-Kyoung, K.: Rotational data augmentation for electroencephalographic data. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 471–474 (2017)

13. Lin, Q., Haoran, Z., Cao, X., Lyu, X., Xu, L., Yang, B., Ou, Y.: Multi-scale feature fusion convolutional neural network for concurrent segmentation of left ventricle and myocardium in cardiac MR images. *J. Med. Imag. Health Inf.* **10**(5), 1023–1032 (2020)
14. Alexander, N., Luc, V.G.: Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR'06), vol. 3. IEEE, pp. 850–855 (2016)
15. Martin, A., Paul, B., Jianmin, C., et al.: TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, pp. 265–283 (2016)

# A Survey of FPGA-Based Deep Learning Acceleration Research



Ziyi Lv  and Jing Zhang 

**Abstract** In a range of fields such as emotion detection, medical image processing and speech recognition, deep learning has recently achieved good results. With the pursuit of more precise results, many scholars try to add more new type network layers to increase the size of the neural network. However, this will lead to deeper and more intricate network models, and training and evaluating models requires intensive CPU calculations and tremendous computing resources which cannot be achieved by general purpose processors. Nowadays, some hardware accelerators such as Field Programmable Gate Array (FPGA) have been employed to accelerate the neural network, and FPGA with reconfigurability and low power consumption are currently applied to improve throughput of deep learning networks at a reasonable price. In this paper, the typical technologies and methods of accelerating deep learning network on FPGA in recent years are reviewed and analyzed with their advantages and disadvantages, and feasible research suggestions for the next research direction are given. It is expected that it will have a certain reference value for researchers in the field of deep learning acceleration and hardware optimization.

**Keywords** Hardware accelerator · FPGA accelerator · Deep learning

## 1 Introduction

In recent years, there has been immense advances in the research of neural networks compared with traditional algorithms. Due to their remarkable results, some neural networks algorithms have emerged, such as Convolutional Neural Network (CNN) [1], Long/short term memory (LSTM) [2], Generative Adversarial Networks (GAN) [3]. Although deep learning algorithms have achieved great success, the training of models requires substantial amounts of data, which are not available in many cases.

---

Z. Lv · J. Zhang ()

School of Information Science and Engineering, University of Jinan, Jinan 250022, China  
e-mail: [ise\\_zhangjing@ujn.edu.cn](mailto:ise_zhangjing@ujn.edu.cn)

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

For example, over 19.6 billion floating point operations (flops) and up to 500 MB model parameters are required by VGG19 for image classification. It can be seen that larger input images will lead to more intricate calculations. Therefore, it is particularly important to accelerating neural networks.

There are mainly two strategies for accelerating neural networks, which are software methods and hardware methods. With the goal of reducing the total calculation of the model without affecting accuracy, software methods roughly consist of weight reduction [4], data quantification and the optimization of algorithm procedure [5]. With the pursuit of reducing the computation of the models, weight reduction is the approximation of the weight matrix using a low-rank matrix. Data quantification is mainly responsible for quantifying weights and neurons in the computing of model to achieve the purpose of reducing the bandwidth and storing requirements. And the optimization of algorithm procedure is primarily to simplify or transform the calculation process pointing to the characteristics of different network models with minimal impact on results. The hardware strategies primarily optimize the structure and layout of the hardware unit for the inherent characteristics of model so that neural network algorithms can be executed quickly while remaining efficient.

Some hardware platforms have been employed to accelerate neural networks. Many researchers choose GPU to accelerate neural networks due to its high memory bandwidth and throughput, but GPU accelerators consume massive amounts of power. Although ASIC accelerators have limited memory bandwidth and computing resources, their power consumption is lower while achieving at least moderate performance compared with GPU accelerators. But the development cycle of ASIC accelerators is relatively long. Compared with ASIC, FPGA is slightly inferior in power consumption, operation under the same design, but its development cycle and flexibility are satisfactory due to its reconfigurability. So good performance and efficiency improvement can be achieved using the FPGA platforms in a shorter time for the same implementation.

## 2 The Focus of Deep Learning Acceleration

Aiming to optimize the network accelerator, the previous research carried out the optimization analysis of the performance mainly based on the roofline model [6] describing the relationship between the theoretical performance of the algorithm and the communication bandwidth, operational intensity under the constraints of computing platform where the X-axis of the roofline model is the operational intensity of the model algorithm, and the Y-axis means the theoretical peak performance of the algorithm.

In the Memory-Bound region (that is, so-called “eave” region): When the operational intensity of the model is less than the upper limit of the operational intensity of the computing platform, the model is in the Memory-Bound region at this time, and the theoretical performance of the model is determined by the upper limit of the bandwidth of the calculation platform (that is, the slope of the eave) and the

operational intensity of the model itself. The larger the bandwidth of the computing platform, that is, the steeper the eave, or the greater the operational intensity of the model, the theoretical performance of the model can increase linearly.

In the Compute-Bound region: When the operational intensity of the model is greater than the upper limit of the operational intensity of the computing platform, the model is in the Compute-Bound area in the current calculation platform, that is, the theoretical performance of the model is limited by the calculation power of the calculation platform and cannot be proportional to the operational intensity. Regardless of the operational intensity of the model, its theoretical performance can only be equal to the calculation power of the calculation platform at most.

According to the roofline model, the accelerator can be optimized from two aspects: one is to improve the communication bandwidth, the other is to improve the operational intensity of the model.

- (1) Improve the communication bandwidth: the bandwidth mentioned in the roofline model is the peak bandwidth of the computing platform. When the actual program runs, data is not transmitted at the peak bandwidth, and the actual bandwidth utilization is very low. Therefore, when the operational intensity is relatively fixed, it is particularly important to improve the bandwidth, and a variety of parallel methods are employed to attain this target, which are mainly divided into data-level parallelism, task-level parallelism and hardware-level parallelism. For data-level parallelism, multi-buffering mechanism is employed to speed up the whole computing time in some research by performing data calculations at the same time during data transmission to cover the time cost of data transmission. For instance, Li et al. [7] employed three on-chip buffers for storing the feature data and streaming it to processor elements alternately. In [8], Two input images are handled on two compute units in FPGA at the same time to exploit coarse-grained data parallelism. Task-level parallelism refers to decomposing a software program into multiple tasks and assigning them to different processors for execution to achieve parallelism, mainly involving the optimization of the software system [9]. For hardware-level parallelism, pipeline technology, often used in some research [7, 10–15], is decomposition of a repeated process into several sub-processes, and parallel execution between sub-procedures, so improving throughput. In addition to some parallel methods, some research such as [5, 14, 16] applies the data quantization strategy.
- (2) Improve the operational intensity: Since the operational intensity is equal to the ratio of the calculation amount and the amount of memory access, reducing the amount of memory access can improve the calculation efficiency of the model when the calculation amount of the model is relatively fixed. In the neural network algorithm, some methods such as loop unrolling and Winograd-based method [17] make use of some common features of the algorithm, such as convolution operation and nonlinear activation function, to increase the data reuse of each calculation. In [18], An efficient convolution acceleration

strategy and dataflow are proposed aimed at minimizing data communication and memory access while maximizing the resource utilization to achieve high performance. Besides, Qiu et al. [5] compressed the weight matrix of the FC layer by using Singular value decomposition (SVD) to reduce memory footprint of the FC layer.

### 3 The Current Status of Deep Learning Acceleration

In this section, the latest technologies of FPGA for accelerating and optimizing network algorithms in deep learning research, such as emotion detection and target detection, are reviewed.

In the field of emotion detection, Hector et al. [11] proposed BioCNN, an EEG-based biological neural network employing FPGA. Pipelining technology is applied to entire the BioCNN module, that is, not only in the convolution operation but also within the max-pooling layer and output module, to minimize logic resources. Using the DEAP dataset, the classification accuracy of BioCNN in valence and arousal is 71.25% and 77.57% respectively and uses  $10 \times - 100 \times$  times less hardware resources than the FPGA based CNN mentioned in [12, 19, 20] under the energy expense of 150 mW. In previous studies, only one hardware classifier based on EEG was reported [21]. Compared with BioCNN, the classifier presented in [21] does not use a pipeline structure and provides hardware accuracy only for valence binary detection while ignoring the arousal dimension. However, BioCNN trades off throughput for more compact architectures in line with the constraints of edge nodes. Due to the need to consider resource utilization, BioCNN uses a more compact architecture, which leads to the reduction of throughput. Of course, BioCNN balances the relationship between resource utilization and throughput as much as possible by hardware parallelization, data partitioning.

In the field of object detection, there are roughly three designs of FPGA accelerators according to the implementation type of convolution algorithm. The first category of schemes directly performs convolution operations by performing numerous MAC operations on a massive number of DSP blocks with the using of CNN parallelism, such as [22, 23]. In [23], Xu et al. implemented a FPGA accelerator based on OpenCL for YOLOv2 under 190 MHz working frequency with a peak throughput of 566GOPS on Arria-10 GX1150 FPGA board. However, the performance of this accelerator is limited by the number of DSP blocks. To solve the resource limitation problem and achieve the requirement, the second type of FPGA accelerator uses low-bit operations, such as XOR or AND, replacing MAC operations demanded by DSP, such as [24, 25]. A pipelined based lightweight YOLOv2 presented at [25] achieves 40.81 frames per second on the Xilinx Inc.zcu102 board. However, there is slight loss of detection accuracy for the trade-off between detection accuracy and hardware resources. Another type of design employs software-level optimization of convolution algorithms, such as sparse convolution schemes [26, 27], or frequency domain convolution schemes [28, 29]. Among them, Wang et al. [27] developed

a FPGA accelerator architecture based on YOLOv2 framework employing a novel sparse convolution algorithm. And the detection accuracy on the PASCAL VOC2007 dataset is 74.45% with a peak throughput of 2.13 TOPS (72.5 fps) under the 211 MHz working frequency on an Intel Arria-10 GX1150FPGA. Comparing with [23], the proposed accelerator improves the detection throughput by  $3.8\times$ . Based on the above analysis, utilizing advanced convolution algorithms is a more promising category among three approaches.

Through the analysis of previous research, it can be found that FPGA based accelerators have many advantages, such as reconfigurability, high parallelism, and high performance with low energy. High parallelism benefits from the editable FPGA logic hardware unit, so that the hardware can be easily optimized using parallelized algorithms. FPGA could be flexibly applied to complex engineering situations thanks to its reconfigurability. In recent research, SparkNet presented in [13] has shown outstanding advantages, effectively compressing convolutional neural networks 150 times. However, FPGA accelerators have some shortcomings. For instance, the reconfigurability of the FPGA brings us certain amounts of time overhead that cannot be ignored.

## 4 Expectation

Through the review of this article, there are several points worthy of further study. First, some current research pays attention to the optimization of the convolution operation, and the optimization of other calculation processes is also worth exploring further.

In addition, there is currently a lack of a more general and user-friendly framework that can optimize the model and evaluate it according to the model and other requirements specified by the user.

## 5 Conclusion

In this paper, the recent developments in the field of FPGA acceleration research by deep learning networks are reviewed, and some advantages and disadvantages of FPGA based accelerators for deep learning are summarized. Finally, some recommendations for enhance the effectiveness of FPGA accelerator are provided.

**Acknowledgements** This research is supported by: (1) 2020-2022 National Natural Science Foundation of China under Grand (Youth) No. 52001039 (2) 2020-2022 Funding of Shandong Natural Science Foundation in China No. ZR2019LZH005.

## References

1. Yann, L., Yoshua, B.: Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **3361**(10) (1995)
2. Xingjian, S., Zhourong, C., Hao, W., et al.: Convolutional LSTM network. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, A Machine Learning Approach for Precipitation Nowcasting, pp. 2672–2680. MIT Press (2015)
3. Ian, J.G., Jean, P.-A., Mehdi, M., et al.: Generative Adversarial Nets. MIT Press (2014)
4. Jiantao, Q., Song, S., Yu, W., et al.: Going deeper with embedded FPGA platform for convolutional neural network. In: Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 26–35. ACM (2016)
5. Chen, Z., Peng, L., Sun, G., et al.: Optimizing FPGA-based accelerator design for deep convolutional neural networks. In: Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 161–170. ACM (2015)
6. Williams, S., Waterman, A., Patterson, D.: Roofline: an insightful visual performance model for multicore architectures. *Commun. Assoc. Comput. Mach.* **52**(4), 65–76 (2009)
7. Xuelei, L., Liangkui, D., Li, W., Fang, C.: FPGA accelerates deep residual learning for image recognition. In: 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, Chengdu (2017)
8. Roberto, D., Griffin, L., Jasmina, V., Paul, C., Graham, T., Shawki, A.: Caffeinated FPGAs: FPGA framework for convolutional neural networks. In: 2016 International Conference on Field-Programmable Technology (FPT). IEEE, Xi'an (2017)
9. Chao, W., Junneng, Z., et al.: Hardware implementation on FPGA for task-level parallel dataflow execution engine. *IEEE Trans. Parall. Distrib. Syst.* **27**(8), 2303–2315 (2016)
10. Chen, Z., Di, W., Jiayu, S., Guangyu, S., Guojie, L., Jason, C.: Energy-efficient CNN implementation on a deeply pipelined FPGA cluster. In: Proceedings of the 2016 International Symposium, pp. 326–331 (2016)
11. Hector A.G., Shahzad, M., Jerald, Y., Ibrahim M.E.: BioCNN: a hardware inference engine for EEG-based emotion detection. *IEEE Access* 140896–140914 (2020)
12. Lei, G., Chao, W., Xi, L., Huaping, C., Xuehai, Z.: MALOC: a fully pipelined FPGA accelerator for convolutional neural networks with all layers mapped on chip. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **37**(11), 2601–2612 (2018)
13. Ming, X., Zunkai, H., Li, T., Hui, W., Victor, C., Yongxin, Z., Songlin, F.: SparkNoC: an energy-efficiency FPGA-based accelerator using optimized lightweight CNN for edge computing. *J. Syst. Archit.* **115**(4), 101991 (2021)
14. Gan, F., Zuyi, H., Song, C., Feng, W.: Energy-efficient and high-throughput FPGA-based accelerator for Convolutional Neural Networks. In: 2016 13th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT). IEEE, Hangzhou (2017)
15. Dong, W., An, J., Xu, K.: PipeCNN: an OpenCL-based FPGA accelerator for large-scale convolution neuron networks. *CoRR* **1611**(02450) (2016)
16. Kaiyuan, G., Lingzhi, S., Jiantao, Q., et al.: Angel-eye: a complete design flow for mapping CNN onto embedded FPGA. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **37**(1), 35–47 (2017)
17. Liqiang, L., Yun, L., Qingcheng, X., Shengen, Y.: Evaluating fast algorithms for convolutional neural networks on FPGAs. In: 2017 IEEE 25th Annual International Symposium on Field-programmable Custom Computing Machines. IEEE, Napa (2017)
18. Yufei, M., Yu, C., Sarma, V., Jae-sun, S.: Optimizing the convolution operation to accelerate deep neural networks on FPGA. *IEEE Trans. Very Large-Scale Integr. (VLSI) Syst.* **26**(7), 1354–1367 (2018)
19. Lei, G., Chao, W., Xi, L., Huangping, C., Xuehai, Z.: A power-efficient and high-performance FPGA accelerator for convolutional neural networks. In: Proceedings of the 12th IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis Companion, pp. 1–2 (2017)

20. Chao, H., Siyu, N., Gengsheng, C.: A layer-based structured design of CNN on FPGA. In: 2017 IEEE 12th International Conference on ASIC (ASICON). IEEE, Guiyang (2017)
21. Waichi, F., Kaiyen, W., Nicolas, F., Yunlun, H., Yude, H.: Development and validation of an EEG-based real-time emotion recognition system using edge AI computing platform with convolutional neural network system-on-chip design. *IEEE J. Emerg. Sel. Top. Circ. Syst.* **9**(4), 645–657 (2019)
22. Jing, M., Chen, L., Zhiyong, G.: Hardware Implementation and optimization of tiny-YOLO network. Springer, Singapore (2017)
23. Ke, X., Xiaoyun, W., Dong, W.: A scalable OpenCL-based FPGA accelerator for YOLOv2. In: 2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). IEEE, San Diego (2019)
24. Hiroki, N., Masayuki, S., Shimpei, S.: A demonstration of FPGA-based you only look once version2 (YOLOv2). In: 2018 28th International Conference on Field Programmable Logic and Applications (FPL). IEEE, Dublin (2018)
25. Hiroki, N., Haruyoshi, Y., Tomoya, F., Shimpei, S.: A lightweight YOLOv2: a binarized CNN with a parallel support vector regression for an FPGA. In: Proceedings of the 2018 ACM/SIGDA International Symposium, pp. 31–40. ACM (2018)
26. Chaoyang, Z., Kejie, H., Shuyuan, Y., Ziqi, Z., Hejia, Z., Haibin, S.: An efficient hardware accelerator for structured sparse convolutional neural networks on FPGAs. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 1–13 (2020)
27. Zixiao, W., Ke, X., Shuaixiao, W., Li, Liu., Lingzhi, L., Dong, W.: Sparse-YOLO: hardware/software co-design of an FPGA accelerator for YOLOv2. *IEEE Access* **8**(99), 116569–116585 (2020)
28. Xianchao, X., Brian, L.: FCLNN: a flexible framework for fast CNN prototyping on FPGA with OpenCL and Caffe. In: 2018 International Conference on Field-Programmable Technology (FPT). IEEE, Naha (2018)
29. Caiwen, D., Shuo, W., Ning, L., Kaidi, X., Yanzhi, W., Yun, L.: REQ-YOLO: a resource-aware, efficient quantization framework for object detection on FPGAs. In: Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 33–42 (2019)

# A Survey of Track Prediction Method of AUV Based on Deep Learning



Yuna Yu , Jing Zhang , and Tianchi Zhang 

**Abstract** Autonomous Underwater Vehicles (AUVs) work in the complex marine environment, and security is the focus of research and application. Track prediction is the key to assist AUV to avoid risk, and it is of great significance to predict the drift track of the AUV in the accident for the later study of fault causes and solutions. The real-time spatial position and ocean environment of AUV are the basic data of track prediction, which are large spatial data with high real-time performance. In this paper, the establishment and development of AUV track prediction model are analyzed based on the characteristics of deep learning that can efficiently process complex data. Secondly, the drift track prediction of AUV with large diving depth after the accident is introduced with the focus of the drift track prediction after the AUV floats to the sea. Finally, the research on the track drifting with the ocean current is prospected to face the situation when AUV floating slowly under the effect of residual buoyancy after the accident of AUV.

**Keywords** Autonomous underwater vehicles · Deep learning · Track prediction

## 1 Introduction

The ocean contains massive amounts of resources, and human exploration of the ocean has never stopped. AUV is the key to explore and develop unknown marine resources and complete underwater intelligent operations. AUV and AI are the products of the same era, and have gone through a long and tortuous development process together. Driven by the contemporary machine learning theory represented by deep

---

Y. Yu · J. Zhang

School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

T. Zhang ()

School of Information Science and Engineer, Chongqing Jiaotong University, Chongqing 400074, China

e-mail: [zhangtianchi@cqjtu.edu.cn](mailto:zhangtianchi@cqjtu.edu.cn)

learning and reinforcement learning, AUV has produced profound changes in perception, control and intelligence. Today, AUV is employed as an unmanned survey platform, and carries the sensor payload according to the pre-programmed trajectory for data collection and sampling in deep sea [1]. However, AUV works in a complex marine environment is still disturbed by many uncertain marine environmental factors. More accurate risk avoidance is an important factor for the development of AUV in the direction of long-range, deep-sea and multi-function. Track prediction is the guarantee for AUV to work safely underwater.

Track prediction was first applied in the aviation field. With the gradual development of shipping, the application of track prediction in the marine field has also been given full attention [2]. In order to ensure the safety of AUV in deep sea, accurate and feasible prediction methods are needed. It is necessary to process and collect the real-time position, navigation attitude, current, tide and other large space data of AUV. The traditional track prediction mostly adopts the method of establishing kinematics or dynamics equations. However, the random influence of high real-time space data leads to the too complex process of establishing equations and low realizability. The advanced data processing and learning ability of deep learning can deal with complex information processing, so it can be used for AUV track prediction.

Human's cognition of marine environment and marine equipment is far from mature, so AUV accident is inevitable. The famous American AUV 'Abe' was wrecked during its underwater exploration mission [3]. In 2013, a Japanese Maritime Self Defense Force AUV used to collect submarine noise and other performance data intelligence was lost [4]. In 2015, China found a malfunctioning AUV of unknown nationality in the South China Sea, which is suspected to be intelligence gathering, and has attracted much attention [5]. In 2016, a serious accident occurred in the sea trial of a deep submergence vehicle in China [6]. After a catastrophic accident of AUV, it is important to study its drift track, predict its water point or bottom point, and salvage it in time, which can provide an important reference for studying the cause of failure and further optimizing the track prediction model. However, there are few researches on the drift track of AUV after the accident, and there are many research results on the drift of the wrecked ship that can be used for reference. Therefore, the drift track prediction after the AUV accident is prospected by analyzing the drift track prediction based on the wrecked ship.

## 2 AUV Track Prediction with Actuation

Track prediction methods are mainly divided into two categories: time prediction and position prediction. Time prediction is to predict the time information when the target reaches the designated position under the condition that the target navigation trajectory is determined and the navigation environment is safe and stable. However, the position prediction is more complex, which is the prediction of the position data of the target after a period of unknown trajectory distance and trajectory time when the target's trajectory is unknown. AUV track prediction is mainly the position

prediction of the target. There are two methods for position prediction, the model-based method and model-free method. Most of the current processing methods with models are realized by adding or modifying the correlation coefficient in the model. Most of the model-free methods are based on data statistics and artificial intelligence.

## 2.1 A Subsection Sample

In the complex marine environment, there are many factors that affect AUV track prediction, such as current, tide. Gao et al. [7] proposed that the current is an important factor affecting the navigation of a submersible, and proposed a method to model the dynamic current environment by using a *B*-spline curve. Considering the anisotropic and time-varying characteristics of current, the path planning algorithm is improved to make use of current and local turbulence, and modify the path in real-time according to the dynamic current information, that is to say, the real-time ocean current information of underwater working target can be obtained, and the track can be accurately avoided and predicted. After that, Sun et al. [8] also studied the AUV route planning method considering the influence of ocean current, modified the level set partial differential equation by combining ocean current field and AUV Navigation speed, and numerically solved the level set equation by using Godunov difference scheme, which evolved and periodically reinitialized the level set function from the narrowband region at the beginning of AUV Navigation. Then the tracking equation is solved by backward iteration to ensure that the AUV can avoid the danger of current field and choose the best navigation route. Recently, Guo et al. [9] continued to study the global path planning method of AUV under the influence of variable current. According to the change of underwater ocean current, Gaussian noise recursive equation is used to estimate the velocity vector state of the next time node of ocean current, so as to ensure that AUV can avoid danger dynamically and adapt to the change of ocean current for track prediction. The establishment of dynamic equation and model method presented in [7–9] are based on the influence of ocean current for the obstacle avoidance and track prediction of the target. However, there are many complex hydrological factors in the marine environment, which have great randomness on the movement of AUV, causing the complex process of equation establishment and unsatisfactory prediction effect.

## 2.2 AUV Track Prediction Based on Deep Learning

Deep learning is an important part of model-free track prediction. Perera et al. [10] proposed an artificial neural network as a mechanism to detect and track multiple targets. Aiming at the problem of navigation target state estimation and heading prediction, an extended Kalman filter algorithm is proposed to estimate and predict its trajectory. The researchers [11] in Harbin Engineering University proposed AUV

track prediction based on BP neural network with the input data of ocean current, tidal current and tide which has a impact on the target navigation. BP neural network is improved by using principal component analysis, adaptive learning efficiency and combination test, and a novel time series is proposed to achieve the accurate prediction of navigation trajectory of AUV. Jilin University proposed an algorithm called DR-N [12], through the sensor and neural network algorithm, the heading angle of AUV is calculated to obtain the three-axis acceleration, and then the target navigation trajectory is calculated.

In addition to the use of deep learning data processing and other features like [10–12], based on the characteristics of deep learning that can approximate the cognition and behavior of humans or animals, and by building a neural network model to learn that humans and animals have inherent self-protection capabilities, AUV can achieve accurate hazard avoidance and track prediction. For example, fuzzy set has excellent thinking and reasoning rules similar to human's 'if–then' mode [13], and can increase the identification ability of the model by combining it with the artificial neural network, so the adaptability of AUV route planning and prediction is improved. In addition, ant colony algorithm [14], the classical bionics algorithm of deep learning [15], refers to imitating the behavior the ant colony will choose the safest and shortest path when returning to the nest after they find food. The algorithm presented in [16] improves the classical ant colony algorithm, and adopts the combination of local pheromone update mode and global pheromone update mode to improve the convergence speed. Combined with the influence factors of a large space seabed environment, the obstacle avoidance and prediction problems are simulated, and outstanding experimental results are obtained.

The traditional process of establishing dynamic equation has achieved remarkable results in AUV track prediction. However, due to the high complexity and randomness of marine environmental factors, the process of establishing dynamic equation is intricate. With the continuous development and optimization of deep learning algorithm, the track modeling method based on deep learning has been widely used. In order to improve the accuracy of track prediction, it is necessary to consider the randomness of complex marine environmental factors by increasing the number of neural network layers. At the same time, continuing the relevant research for the slow convergence speed of depth neural network has important research significance.

### 3 AUV Track Prediction Without Actuation

At present, the research of AUV safety technology mostly focuses on fault control and safety design [17], and has achieved adjective research results [17], which ensures the safety of AUV during navigation. However, AUV accidents often occur in complex marine environment. The AUV will drift laterally under the action of current and vertically upward under the action of residual buoyancy, and then float out of the sea to continue to drift.

### 3.1 Sea Surface Drift Track Prediction of AUV

The AUV without actuation drifts in the sea for a period of time, then breaks through the thermosphere under the action of residual buoyancy, and floats out of the sea to continue to drift. Although the drift of marine objects such as ships is different from AUV, the theory and method can be used for Ref. [18], which is called the prediction of drift track of sea surface objects. The drift of the target objects on the sea is mainly affected by wind speed and current velocity. In view of the influence of wind factors, Gu et al. [19] established a motion model to simulate the drift of the target based on the influence of wind drift and wind induced deflection on the ship movement. After that, Allen [20] established a wind drift model through a large number of marine test results for the first time. Based on the wind-induced drift model, Xiao et al. [21] studied and analyzed the trajectories of drifting objects on the sea, established a prediction and simulation system for the trajectories of drifting objects at the estuary of the Yangtze River and its adjacent seas, and predicted the trajectories of drifting objects on the sea surface. Yuan et al. [22] aim at the floating buoy drift problem based on Monte Carlo method, the system of drift tracking and prediction of buoys can be used on Web pages by WebGIS. The system can automatically access the environment data, and use the SARMAP model to simulate the drift according to the buoy parameter information to predict the drift trajectory and the optimal search area of the buoy. All of the above are based on the influence of wind or the establishment and combination of wind-induced drift model to establish the simulation system of sea surface target drift track prediction.

In addition, Zhou et al. [23] based on Bayesian method, constructing a Max-Max programming model to solve the optimal search strategy with the aim of discovery probability. The disadvantage is that the real-time and intelligence of the next search task can be corrected and adjusted in time. Then, Zhang [24] proposes a model based on probability to study the trajectory prediction of the nondynamic drift under the action of current and wind, which can continuously predict the velocity and position of the object at any time. The downside is that the real environment is much more complex, and the uncertainty of the location of the prediction target will increase. Miron et al. [25] use the ground drift data tracked by satellites in Indian Ocean History to proposes a Markov chain model based on probability problem to represent the ocean drift path of debris on the missing mh370 aircraft of Ma Hang. However, the framework based on probability algorithm cannot solve the problem of data assimilation of many factors. The above method based on probability algorithm and mathematical model can predict the drift track of the target with many sea environmental factors, but the disadvantages are obvious. When the influence factors are increasing, the accuracy of the prediction results will decrease.

The drift of sea surface target is mainly caused by wind, and the draft by current of AUV also needs to be considered. Ocean circulation and other data information are generally historical data, while wind-induced drift and other data are real-time data. Accurate prediction of sea surface target drift requires the combination of historical data and real-time data. In addition, it is still insufficient to solve the

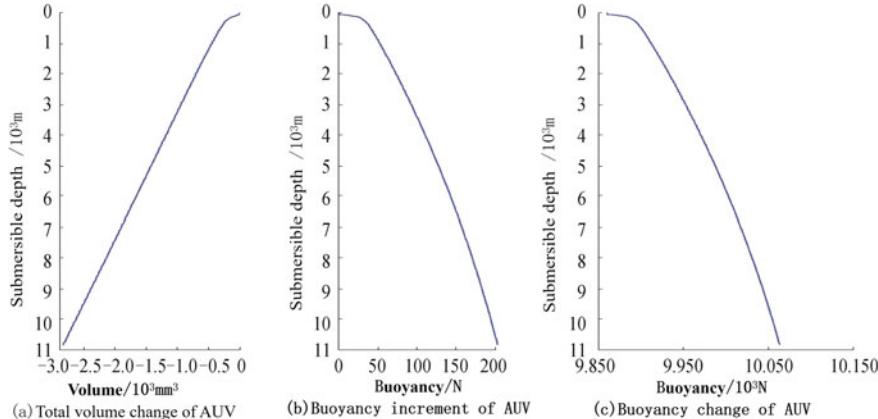
problem of appropriate data assimilation of various influencing factors, and the real-time performance of timely correction and adjustment of the next search task is poor. The geometric structure of sea surface drifts is not completely regular, and the prediction error of wind pressure angle of irregular drifts is still relatively large. At present, there are very little literature on AUV floating in the sea and sea surface drift, which needs to be further explored.

### **3.2 Drift and Buoyancy Variation in the Sea of AUV Without Actuation**

The drift after the loss of control in AUV water with large potential is mainly the nondynamic lateral drift under the action of current and longitudinal upward drift under the action of residual buoyancy. In the process of drift, the environmental physical parameters change significantly, and the gravity and buoyancy state change accordingly. This change is directly related to whether the AUV can break through the thermosphere and float to the sea surface under the residual buoyancy, so it must be considered. At present, the relevant literature is aimed at the drift track of normal ships or AUV in the sea with driving force. There is no relevant literature about the prediction of AUV track in the sea due to the state of no dynamic drift after the crash, so it needs to be explored.

Liu et al. [26] measure buoyancy based on propeller and acceleration calculation. According to the density change under the water depth of 10,000 m and uncertainty of mechanical parameters of new materials, a real-time buoyancy measurement method of a million meters underwater robot is proposed. According to the variation of the sea density, gravity acceleration and robot volume, Lu et al. [27] can calculate the variation of the buoyancy state of the robot according to the ARV of the whole sea depth. The calculation method of the buoyancy balance of the ARV in the whole sea is obtained by calculating and combining the ARV deep submergence data of ‘Haidou’ and the experimental data of compensating oil compression. The results show that the ballast required for buoyancy balancing at 11,000 m depth is 8.9 kg, and the sea test is carried out. This method is of great significance for the realization of one dive. Based on the uncertainty factors in deep water environments such as seawater temperature, salinity, pressure and density, as well as the analysis and Research on the earth gravity acceleration and its abnormal value sum, Jiang and Li [28] have combined the volume change law of components of different structures and materials under the change of ambient temperature and pressure in the test prototype, and the buoyancy curve of the robot in the process of submarine floating shown in Fig. 1.

The results show that the effect of gravity anomaly on the resultant force of abyss profile is no more than 1 N. For the AUV with tonnage and composed of buoyancy material and glass spherical shell, the buoyancy will increase by 204 N when the vehicle reaches 11 km. The results provide accurate prediction for the static equilibrium changes of the robot in the process of submarine floating in the



**Fig. 1** Change of volume and buoyancy of the FOD-AUV with depth

deep sea. Wang et al. [29] analyzed the changes of heavy buoyancy in the process of submarine and up floating of deep AUV in the whole sea. Based on the measured deep water physical parameters, the sea temperature is about 28.4 °C. Between 0 and 300 m, the temperature decreased sharply with the increase of depth, and then the decreasing trend was slow. The temperature from 2000 to 11,000 m is about 2–2.4 °C. The density of seawater is about 1022.0 kg/m $^3$  at the sea surface, which increases sharply between 0 and 300 m, and then the increasing trend slows, which is positively proportional to the depth. At 10,860 m, the density is about 1074.5 kg/m $^3$ .

The research on the variation law of buoyancy of AUV with large potential depth is mainly in the past ten years, in the deep marine environment, the salinity, temperature, sea water specific gravity that affect the buoyancy of AUV are all changing. The law of buoyancy changes in the layer over temperature is more complex. The slightly worse residual buoyancy will make it difficult for AUV to float out of the sea under the temperature layer. Therefore, it is of great significance to study the law of buoyancy change and the minimum residual buoyancy of AUV in deep potential.

## 4 Conclusion

This paper introduces the track prediction method of AUV in two states of normal driving force and loss of driving force after an accident. By comparing with the method of establishing dynamic equation, this paper introduces the track prediction method based on deep learning. Secondly, the drift mode of AUV without driving force is introduced, and the drift track prediction method of sea surface target is analyzed by learning from the drift track prediction method of wrecked ship. Finally, the prediction method of AUV drift track in the sea is prospected. There are many

differences between the drift of AUV in the sea and the drift of target on the sea, and there are little related literature, so the related research needs to be continued.

**Acknowledgements** This research is supported by (1) 2020-2022 National Natural Science Foundation of China under Grand (Youth) No. 52001039 (2) 2020-2022 Funding of Shandong Natural Science Foundation in China No. ZR2019LZH005.

## References

1. China AI 2.0 Development Strategy Research Project Group: China AI 2.0 Development Strategy Research. Zhejiang University Press, Hangzhou (2018)
2. Yuke, H., Wei, X., Xiaoxuan, H., et al.: Ship track prediction based on recurrent neural network. *Syst. Eng. Electron. Technol.* **042**(004), 871–877 (2020)
3. Xianbo, X., Caoyang, Y., Qin, Z.: On intelligent risk analysis and critical decision of underwater robotic vehicle. *Ocean Eng.* **140**, 453–465 (2017)
4. Xisheng, F., Yiping, L., Hongli, X.: Next generation marine robot. *Robot* **33**(1), 113–118 (2011)
5. Japan's maritime self defense force lost an unmanned submersible device. <http://asahichinese.com/article/society/AJ201401290045>. Asahi Shimbun, Last accessed 2014
6. Mario, B., Gwyn, G., James, F.: A behavioral probabilistic risk assessment framework for managing autonomous underwater vehicle deployments. *J. Atmos. Oceanic Tech.* **29**(11), 1689–1703 (2012)
7. Bo, G., Demin, X., Fubin, Z., Weisheng, Y.: Ocean current modeling and its application in path planning. *J. Syst. Simul.* **22**(4), 957–961 (2010)
8. Tianlong, S.: Research on AUV Route Planning Method Considering Current Influence. Harbin Engineering University, pp. 35–46 (2016)
9. Xinghai, G., Mingjun, J., Weidan, Z., Jinwei, Z., Lingrui, K.: Improved QPSO algorithm for dynamic path planning of autonomous underwater vehicle in variable current environment. In: System Engineering Theory and Practice (2020)
10. Lokukaluge, P.P., Paulo, O.C., Guedes, S.: Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction. *IEEE Trans. Intell. Transp. Syst.* **13**(3), 1188–1200 (2012)
11. Chun, Y.: Research on AUV Track Prediction Method Based on BP Neural Network. Harbin Engineering University (2014)
12. Yanxin, X.: Research on Dead Reckoning Algorithm of AUV Based on Neural Network. Jilin University
13. Zhouzhou, L.: AUV path planning based on fuzzy neural network. *Microprocessor*, 43–45 (2015)
14. Hui, W., Xiaoyang, H.: Research on UAV path planning based on ant colony algorithm. *Sci. Tech. Inf.* **583**(10), 35–36 (2020)
15. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. *IEEE Comput. Intell. Mag.* **1**(4), 28–39 (2006)
16. Nannan, Z.: Research on Path Planning and Tracking Method of Underwater Vehicle. University of Science and Technology, Jiangsu (2019)
17. Yousheng, W., Yiyu, Z., Shuyan, L., Chuanrong, W.: Research on technology development of intelligent unmanned underwater vehicle. *China Eng. Sci.* **22**(6), 026–031 (2020)
18. Xi, Z., Yixun, Z., Jie, Z.: An algorithm for sea ice drift retrieval based on trend of ice drift constraints from sentinel-1 SAR data. *J. Coastal Res.* **102**(sp1), 113–126 (2020)
19. Wenxian, G.: Wind induced drift and wind induced deflection. *World Shipping*, pp. 45–46 (1995)

20. Allen, A.A., Plourde, J.V.: Review of leeway: field experiments and Implementation. Review of leeway field experiments and implementation, p 351 (1999)
21. Wenjun, X., Panjun, D., Maoli, G.: Research and application of Shanghai coastal search and rescue prediction model system. *Ocean Forecast* **30**(4), 81–88 (2013)
22. Li, Y., Hong, R., Peiyi, L.: Research on the prediction and analysis system of buoy drift trajectory. *Navigation*, pp. 24–28 (2020)
23. Changyin, Z., Yutang, Z., Yaxing, S.: Probability model of aircraft crash detection based on Bayesian information updating. *Math. Model. Appl.* **4**(2), 71–78 (2015)
24. Jinfen, Z., Ângelo, P., Teixeira, C., Guedes, S., Xinping, Y.: Probabilistic modelling of the drifting trajectory of an object under the effect of wind and current for maritime search and rescue. *Ocean Eng.* **129**(1), 253–264 (2017)
25. Miron, P., Beron-Vera, F.J., Olascoaga, M.J., Koltaï, P.: Markov-chain-inspired search for MH370. *Chaos* **29**(4), 041105 (2019)
26. Xinyu, L., Yiping, L., Xisheng, F.: Real time buoyancy measurement method for 10000 meter underwater vehicle. *Robot* **40**(2), 216–221 (2018)
27. Yang, L., Yuangui, T., Jian, W., Cong, C., Xingya, Y.: Calculation method for buoyancy trim of full sea deep ARV. *Robot* (2020)
28. Yanqing, J., Ye, L., et al.: Calculation of gravity and buoyancy of underwater robot. *J. Harbin Eng. Univ.* **41**(04), 481–486 (2020)
29. Youkang, W.: Simulation Research on Submerged Floating Motion of Full Sea Deep AUV. Harbin Engineering University, pp. 29–44 (2019)

# Recognition of Farmers' Working Based on HC-LSTM Model



Wenxin Zhao , Jinpo Xu , Xiang Li , Zhaoqi Chen , and Xin Chen

**Abstract** The standardization of farmer's working behavior greatly affects the quality of agricultural products. The records of farmers' work should be included in the agricultural product traceability system, but it is not covered in the existing traceability system. Besides, most of the data is directly stored as image data, which occupies too much storage space, and the effective information cannot be retrieved, therefore, made it necessary to recognize farmers' working behavior automatically. In this paper, farmer's labor behavior recognition has been formulated as a spatiotemporal video classification problem, and an end-to-end trainable model specifically was designed to achieve the recognition of farmers' working behavior. The Hierarchical Convolutional LSTM neural network (HC-LSTM) was modified. HC-LSTM outperforms 3DCNN by 3.0% on FLD, a dataset containing 577 short videos involving 4 typical farming behaviors: spraying pesticides, hoeing the ground, weeding, and planting seedlings.

**Keywords** Farmers' working · HC-LSTM · ConvLSTM

## 1 Introduction

The paramount reason for the existing food safety problem of agricultural products is the incomplete, opaque, and asymmetry of production information [1, 2]. In order to face this problem, it deems to be a critical aspect to establish a reasonable and reliable traceability system for agricultural products [3]. Only when the information of every link in the agricultural product supply chain is true, complete and transparent, can consumers be aware of the cultivation of agricultural products for details.

The working process of farmers will have a relatively large impact on the quality of agricultural products. On the one hand, inaccurate cultivation will reduce the quality of agricultural products. On the other hand, non-standard operations, such as spraying, will affect pesticide residues in agricultural products and cause food

---

W. Zhao · J. Xu · X. Li · Z. Chen · X. Chen (✉)  
China Agricultural University, Beijing, China  
e-mail: [chxin@cau.edu.cn](mailto:chxin@cau.edu.cn)

safety issues. Thus, the farmer's working behavior is indispensable information in the traceability of agricultural products. However, the authenticity of existing working records is questionable which are mainly filled out manually by personnel.

With cameras and other digital equipment gradually used in farmland scenes, it becomes possible to record and analyze farmers' working behavior through video. Therefore, this paper explores the use of deep learning models to analyze video recordings, and then realizes automatic recognition of farmers' labor behaviors, and provides effective farmers' working information for the traceability system.

In essence, recognition of farmers' labor behavior is a spatiotemporal video classification problem. According to the recent advances in deep learning, various pre-trained convolutional network (ConvNet) models [4] are made available for extracting image features. Since the introduction of AlexNet [5], 2D CNNs have made a significant achievement in the field of static images recognition. However, 2D CNNs are unable to model temporal information and motion patterns, which are not suitable for video analysis. Recently, the LSTM-RNN networks have been successfully employed for modeling temporal dynamics in videos. In [6–9], some useful insights are provided on how to tackle this problem. Disappointingly, most of the aforementioned deep learning methods treat video as a frame/optical flow image sequence for video representation learning, while the temporal evolution across consecutive frames is not fully utilized [10]. To deal with this issue, 3D CNN proposed by Ji et al. [11] is one of the earlier works to directly learn the spatiotemporal representation of a short video clip. In 2015, Tran et al. [12] proposed 3D Convolutional Networks (C3D), which used 3D convolution kernels to model video time information and achieved an accuracy of 0.852 on Sport1M [13]. In addition, the ConvLSTM network proposed by Shan et al. extends FC-LSTM to ConvLSTM to better capture the spatiotemporal correlations.

The main contribution of this work is the proposal of Hierarchical Convolutional LSTM neural network (HC-LSTM). HC-LSTM is based on ConvLSTM, a classifier formed by stacking multiple ConvLSTM layers, also as an end-to-end trainable model for farmer labor recognition. In order to increase the effectiveness of the experiment, the C3D model is used for comparison. In the end, the various methods of this paper are applied to the video analysis of farmers' labor behavior recognition.

## 2 Materials and Methods

### 2.1 FLD

**Data Acquisition.** In the past ten years, although many large-scale video datasets on action recognition have emerged, there is no video related to recognition of farmers' labor behavior, which cannot be used as the dataset. In this research, our dataset called FLD, collected from public databases such as Bilibili, Tencent Video, Youku, and consisted of 53 videos related to farmers' actions, which included four the most

common and classic categories: Spraying pesticides, Hoeing the ground, Weeding and Planting seedlings with the frame rate of the video is around 25 fps. The resulting dataset is shown in Fig. 1.

**Data Pre-processing.** The steps of data pre-processing are divided into the following:

- First, we convert the video into an image sequence, and then arrange and name the pictures according to the order in which they appear in the video.
- In the second step, we divide these images into segments in sequence, where each segment contains 50 frames.
- In the third step, we manually remove the segments where there are no farmers or the scene switching is too fast.
- In the fourth step, we use interval sampling to extract the images in the segment as samples, with 5 frames as the interval unit.
- Finally, we use a semi-automated method to extract the area where farmers work in agriculture. Specifically, the yolov3 [14] model is used as a human body detector to locate the movement area of farmers contained in each image in the sample sets. Since multiple people are found in one picture, the area of each human body in the first picture in the sample sets is used as the benchmark, and the nearest neighbor algorithm (KNN) [15] is used as the person association matching algorithm.

After the above steps, the FLD data set is obtained. In the FLD, there are a total of 53 videos, each type of action has 13–14 videos, divided into 119–163 samples, as shown in Table 1.



**Fig. 1** The FLD samples

**Table 1** FAD statistics

Classification	Number of videos	Number of samples
Spraying pesticides	14	119
Hoeing the ground	13	146
Weeding	13	163
Planting seedlings	13	149

## 2.2 Models

**HC-LSTM.** In this section, we present our HC-LSTM model which is based on ConvLSTM. The classic LSTM [16] structure expands the data in one dimension for prediction, which can better solve the time correlation, but (Fully connected LSTM, FC-LSTM) [17] can only extract time-series information, not spatial features. Shi [10] et al. proposed ConvLSTM in 2015 by extending the FC-LSTM to have a convolution structure in both input-to-state and state-to-state transitions. ConvLSTM layer retains the advantages of FC-LSTM and applies to spatiotemporal data due to its inherent convolution structure. In this way, not only can the temporal correlation be obtained, but also spatial features can be extracted like a convolutional layer. The working principle of the ConvLSTM model could be generally given by:

$$f^{(t)} = \sigma[W^{(f)} * x^{(t)} + U^{(f)} * h^{(t-1)} + b_f] \quad (1)$$

$$i^{(t)} = \sigma[W^{(i)} * x^{(t)} + U^{(i)} * h^{(t-1)} + b_i] \quad (2)$$

$$o^{(t)} = \sigma[W^{(o)} * x^{(t)} + U^{(o)} * h^{(t-1)} + b_o] \quad (3)$$

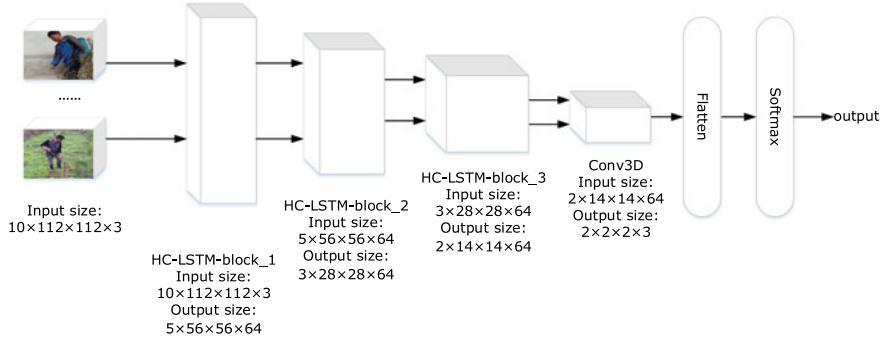
$$\tilde{c}^{(t)} = \tanh[W^{(c)} * x^{(t)} + U^{(c)} * h^{(t-1)} + b_c] \quad (4)$$

$$c^{(t)} = \sigma[f^{(c)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)}] \quad (5)$$

$$h^{(t)} = o^{(t)} \circ \tanh(c^{(t)}) \quad (6)$$

$W^{(f)}$ ,  $W^{(i)}$ ,  $W^{(o)}$  represent the weights of the forget gate, input gate, and output gate at time t.  $U^{(f)}$ ,  $U^{(i)}$ ,  $U^{(o)}$  indicate the hidden weights of the forget gate, input gate, and output gate at t-1, and  $b_f$ ,  $b_i$ ,  $b_o$  denote the biases. where ‘ $\sigma$ ’ denotes sigmoid function is an activation function, so as  $\tanh$ , and ‘ $\circ$ ’, as same as the original LSTM equation, denotes the Hadamard product. The difference is that all inputs  $x^{(1)}, \dots, x^{(t)}$  and memory units  $c^{(1)}, \dots, c^{(t)}$ , hidden states  $h^{(1)}, \dots, h^{(t)}$  and input gates, forget gates and output gates are all three-dimensional tensors. In addition, ‘ $*$ ’ denotes the convolution operator. Based on the above-mentioned ConvLSTM, a recognition model of farmers’ labor behavior is designed, which is called the HC-LSTM model. The framework of the model is shown in Fig. 2.

The input data of HC-LSTM adopts the short-interval format, taking 10 112 × 112 RGB pictures with skipped frames as the model input. The sampling fps is about 25, in other words, 1 frame is extracted every 0.04 s. Compared with the traditional continuous input images, the short-interval input can more clearly capture the characteristics of the data in time and space, making the convolutional layer easy to extract the features of the input sequence images.

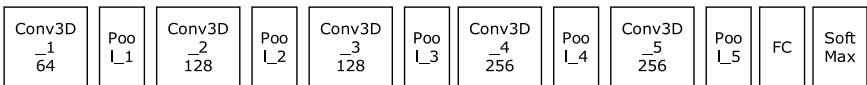


**Fig. 2** HC-LSTM architecture

The network includes 3 HC-LSTM-blocks, 1 C3D-block and 3 fully connected layers, and the final output is the classification result. The size of the convolution kernel of the ConvlSTM layer is  $3 \times 3$ , and the number of convolution kernels is set to 64. All of the pooling layers are  $2 \times 2 \times 2$ , perform down-sampling processing on the information extracted by the convolutional layers. In order to prevent overfitting, add a dropout layer to HC\_LSTM\_block\_3. Convolutional filters in C3D\_block have the size of  $3 \times 3 \times 3$ . Since the RGB image has 3 channels, the number of convolution filters is set to 3. Then the pooling layer with a pooling size of  $1 \times 7 \times 7$  is formed. Finally, after 3 fully connected layers, the final feature vector is calculated to achieve classification.

**C3D Baseline.** 3D ConvNets (C3D), a classical and effective spatiotemporal feature extractor, is used as a baseline for comparison experiments. Proposed by Ji et al. [11], a 3D convolutional network can be applied to various fields such as action recognition and video classification. Tran et al. [12] demonstrate that C3D extraction of spatiotemporal features advantages of good generalization, high calculation efficiency, and simple calculation. The C3D structure designed in this paper is shown in Fig. 3.

In Fig. 3, shows a schematic diagram of C3D. The size of all convolution kernels is  $3 \times 3 \times 3$ .



**Fig. 3** C3D network structure

**Table 2** Model parameters

Category	Evaluation index
Optimizer	Adam optimizer
Measure to prevent overfitting	Data augmentation and dropout
Learning rate	Exponential decay

### 2.3 Experimental Environment and Parameters Design

The experimental environment of the research is i9-10900 K CPU and a 3080rtx graphics card. In the training stage, we use mini-batch and fivefold cross-validation training. In particular, as shown in Table 2, the network parameters are optimized by Adam and the initial learning rate is set as 0.001. In terms of data augmentation, each sample has a probability of 0.5 to be horizontally flipped during training. Moreover, dropout is used to reduce the effect of over-fitting. To avoid an excessive learning rate that might make the network difficult to converge and cause the weight to linger at the optimal value, the learning rate is used to decay with a decay of 0.1 every 10 training epochs. In the experiment, the model was trained for 30 epochs, and the data was scrambled again every 5 epochs.

The performance of the models is evaluated by measuring video classification average precision (*AP*), *recall* and *F1* score on the test set. *AP* indicates the proportion of true positives among the cases that are predicted to be positive, which could be expressed as  $AP = \frac{TP}{TP+FP}$ . *recall* denotes the proportion of all instances that are truly positive that are predicted to be positive, which could be generally given by  $recall = \frac{TP}{TP+FN}$ . *F1* as the harmonic average of *AP* and *recall*, could be generally given by  $F1 = \frac{2 \cdot AP \cdot recall}{AP + recall}$ .

*TP* (True Positives) is predicted to be the number of positive samples that are actually positive samples, *FP* (False Positives) is predicted to be the number of positive samples that are not actually positive samples, and *FN* (False Negatives) is the number of positive samples that are predicted to be negative samples.

## 3 Results and Discussion

### 3.1 Exploring the Hyper-Parameters in HC-LSTM

The difference in the number of convolution filters will have a great impact on the performance of the model. Hence, in order to determine the number of convolution filters in each block, the experiment will adjust the number of convolution filters in each block. The number of kernels is set to m, 2 m, 3 m, 4 m, 5 m, respectively. The experimental results of HC-LSTM are reported in Table 3.

Table 3 presents the performance of the model varies according to the number of convolution kernels. When the number of convolution kernels is 4 m, F1 reaches

**Table 3** Experimental results produced by different convolution kernels

Filters	mAP	mRecall	F1	Time (s)	Total parameters
<i>m</i>	0.8434	0.8379	0.8406	452.59	468,335
<i>2 m</i>	0.9347	0.9259	0.9303	578.22	1,027,727
<i>3 m</i>	0.8992	0.8939	0.8965	716.52	1,955,759
<i>4 m</i>	0.9344	0.9303	0.9324	872.30	3,252,431
<i>5 m</i>	0.8924	0.8817	0.8870	926.97	4,398,063

a maximum of 0.9324, mAP (mean AP) and mRecall (mean Recall) is 0.9344 and 0.9303 respectively, and the training time is 872.30 s. When the number of convolution kernels is 5 m, F1 is 0.8870, which is 4.86% lower than the case of 4 m. This indicates that if the number of convolution kernels is appropriately increased, the model will fit better. If the number of convolution kernels is set too large, the training time will increase, but the effect will decrease, and so is the efficiency.

When the number of convolution kernels is 2 m, F1 reaches 0.9303, with a training time of 578.22 s, which is only 0.22% lower than the maximum F1; but the training time and total parameters are reduced by 33.71% and 68.4%, respectively. Under comprehensive consideration, the model performs best when the number of convolution kernels is set to 2 m, so the number of convolution kernels is determined to be 2 m.

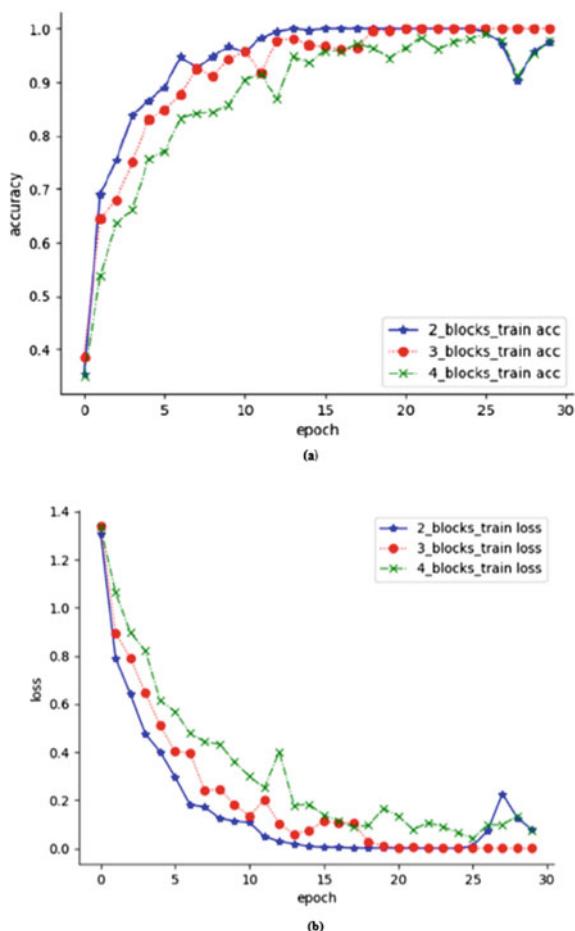
The experiment continues to regulate the number of blocks and dropout rate in HC-LSTM. The accuracy and loss of each model vary with the number of training epochs as displayed in Fig. 4.

In Fig. 4, the red line represents the loss and accuracy of HC-LSTM (3 blocks) in 30 epochs; the blue line signifies HC-LSTM (2 blocks), and the green line denotes HC-LSTM (4 blocks). After 10 epochs, the model begins to converge, and the accuracy gradually rises. HC-LSTM (2 blocks) and HC-LSTM (3 blocks) perform better than HC-LSTM (4 blocks). Although HC-LSTM (2 blocks) converges the fastest during training, it is unstable in the late training period, and its performance on the test set is not satisfactory.

In order to better detect the model ability, a mechanism to prevent over-fitting was adopted in the following experiment. A comparative experiment was performed on HC-LSTM (3 blocks) with dropout rate 0.3, dropout rate 0.5, and dropout rate 0.8.

We report the results of each model with different blocks and dropout rates in Table 4. HC-LSTM (3 blocks and dropout = 0.5) achieves the highest F1 of 0.9339, the mAP and mRecall are 0.9315 and 0.9364 respectively. Followed by HC-LSTM (3 blocks and dropout = 0.3) F1 boosts 0.9303, mAP 0.9347 and mRecall 0.9259. The F1 of the 4 blocks and 2 blocks models are 0.0865 and 0.0252 lower than 3 blocks and dropout = 0.5, respectively. That suggests the ability of HC-LSTM (3 blocks) is significantly better than other models. HC-LSTM (3 blocks and dropout = 0.5) performs best among the nets described previously.

**Fig. 4** The number of blocks search



**Table 4** Performance comparisons in terms of mean AP, mean Recall, F1 score etc

Method	mAP	mRecall	F1	Time (s)	Parameters
HC-LSTM (2 blocks)	0.9111	0.9063	0.9087	512.06	793,987
HC-LSTM (3 blocks and dropout = 0.3)	0.9347	0.9259	0.9303	578.22	1,027,727
HC-LSTM (3 blocks and dropout = 0.5)	0.9315	0.9364	0.9339	582.67	1,027,727
HC-LSTM (3blocks and dropout = 0.8)	0.8657	0.8539	0.8598	563.35	1,027,727
HC-LSTM (4 blocks)	0.8476	0.8472	0.8474	603.88	1,312,151

**Table 5** Comparisons of HC-LSTM, C3D in terms of mAP, mRecall, F1, Time and total parameters on FFAD

Classification	HC-LSTM (3 blocks and dropout = 0.5)			C3D		
	AP	Recall	F1	AP	Recall	F1
Spraying pesticide	0.9166	0.9565	0.9361	0.84	0.9130	0.8749
Hoeing the ground	0.9032	0.9655	0.9333	0.9	0.9310	0.9152
Weeding	0.9355	0.8529	0.8923	0.9090	0.8824	0.8955
Planting seedlings	0.9705	0.9705	0.9705	0.9686	0.9117	0.9393
Mean value	0.9315	0.9364	0.9339	0.9044	0.9059	0.9051
Time (s)	582.67			288.58		
Parameters	1,027,727			3,352,536		

### 3.2 Comparative Experiments

Table 5 shows the performance and time efficiency of HC-LSTM and C3D on FLD. HC-LSTM model boosts higher than C3D in the average value of each evaluation index. Among them, the F1, mAP, and mRecall of HC-LSTM are 3.18%, 3.0%, and 3.37% higher than that of C3D, respectively. Besides, the scale of HC-LSTM has been reduced a lot for the parameters of HC-LSTM being 69.34% lower than C3D. From the perspective of generalization ability, these results basically prove the advantages of HC-LSTM in exploring spatiotemporal information.

## 4 Conclusions

This article focuses on identifying effective farming behavior categories from farmers' working videos. First, we collected the FLD dataset and proposed an end-to-end model based on the HC-LSTM, successfully bring about effective recognition of farmers' labor behavior. The experimental comparison demonstrates HC-LSTM outperforms C3D. From the experimental results, the F1 of HC-LSTM is 3.18% higher than that of C3D, and the scale of the model is reduced by 69.34%.

The future work is as follows: First, we will enrich the dataset in quantity and type. Only a large-scale dataset can more effectively verify the effect of the model and apply the model to practice. Second, further explore the structure of the model, including improving the overfitting mechanism, strengthening generalization ability, and extending the model to the field of farming behavior detection. Behavior recognition of farmers' working alone cannot be satisfied with practical application. Only by further behavior detection could it be used in life.

## References

1. Zhai, H.Y.: Research on the application of block chain technology in the traceability of agricultural products safety information. *Int. Things Technol.* **010**, 90–92 (2020)
2. Liang, K., Shen, M., Ge, Y., Lu, S.: Acquisition and transmission system for grain traceability based on two-dimensional barcode and ARM. *Trans. Chin. Soc. Agric. Eng.* **28**, 167–171 (2012)
3. Elise, G., Barry, K., Fred, K., Linda, C., Kenneth, N., Gregory, P.: Traceability in the U.S. Food Supply: Economic Theory and Industry Studies. *Agricultural Economics Reports* (2004)
4. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. ACM (2014)
5. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: *NIPS* (2012)
6. Ng, Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Toderici, G.: Beyond short snippets: deep networks for video classification. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)
7. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* 3104–3112 (2014)
8. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 352–364 (2018)
9. Srivastava, N., Mansimov, E., Salakhutdinov, R.: Unsupervised learning of video representations using LSTMs. *JMLR.org* (2015)
10. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. MIT Press (2015)
11. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013)
12. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: *IEEE International Conference on Computer Vision* (2015)
13. Andrej, K., George, T., Sanketh, S., Thomas, L., Rahul, S., Feifei, L.: Large-scale video classification with convolutional neural networks. *Comput. Vis. Pattern Recogn.* (2014)
14. Joseph, R., Ali, F.: YOLOv3: an incremental improvement. arXiv e-prints, pp. 89–95 (2018)
15. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithms. *Mach. Learn.* **38**, 257–286 (2000)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–80 (1997)
17. Graves, A.: Generating sequences with recurrent neural networks. *Comput. Sci.* (2013)

# Recognition of Corn Diseases and Insect Pests Based on Residual Network and Transfer Learning



Chun Liao , Jiahao Wang , Qilin Xiong , and Wanlin Gao

**Abstract** The recognition of crop diseases and insect pests based on deep learning has the characteristics of high accuracy and fast speed. This paper establishes a residual network ResNet50 based on transfer learning to identify corn diseases and insect pests. Firstly, the original data is rotated and flipped horizontally at random after being randomly resized and cropped to 256 \* 256 specifications. Then in image center it is cut into 224 × 224 and converted to Tensor and normalization and other data enhancement operations. Finally, the parameters of the pre-trained model based on flower classification is transferred to the new model for training. At the same time, this article also compares the accuracy of VGG16 and ResNet50 without transfer learning on this data set, and concludes that the prediction accuracy of the residual network ResNet50 based on transfer learning is as high as 96.42%, which illustrates that the high efficiency of the model has high efficiency in the identification of corn diseases and insect pests.

**Keywords** Residual network · Transfer learning · Corn diseases · Insect pests

## 1 Introduction

As an important food crop, corn is widely distributed all over the world. At that time, the corn output in China has exceeded 250 million tons since 2018, and its planting area ranks second after rice, as a large corn planting and consumer country [1]. In addition to edible methods, corn is also widely used in animal husbandry feed, medicine and light industry and the output of corn can promote the development of the national economy.

---

C. Liao · J. Wang · Q. Xiong · W. Gao ()

Key Laboratory of Agricultural Information Standardization, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China  
e-mail: [gaowl@cau.edu.cn](mailto:gaowl@cau.edu.cn)

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

The current diseases affecting corn yield mainly include gray spot [2], leaf spot, maize mosaic virus, rust and so on. Through inspection, if corn disease can be found in time, it will have a huge boost to corn planting. In the past, the inspection of corn pests and diseases was mainly manual, but the manual inspection method not only consumes a lot of time but also causes a certain probability of misjudgment, which severely restricts the efficiency of corn pests and diseases detection [3]. With the development of deep learning, more and more scholars have begun to focus on using deep learning to identify pests and diseases. Using lightweight convolutional neural network for image recognition, not only the recognition speed is fast, but the recognition accuracy is also high [4]. Therefore, the use of computer vision technology to identify corn diseases and insect pests studied in this paper has certain research value in the field of corn planting. At the same time, this paper uses data enhancement operations such as random rotation of the original data, random horizontal flip, and cropping of the center of the picture to 224 \* 224, so that the performance of the model is more superior. The pre-trained flower classification model parameters are transferred and learned into the new model, and finally identified through the residual network to make the model training speed and effect more perfect.

## 2 Research Status at Home and Abroad

With the development of deep neural networks, computer vision technology has been rapidly improved. Jia Shaopeng and others have conducted research on deep belief networks (DBN) [5], convolutional neural networks (CNN), recurrent neural networks (RNN), generative adversarial networks (GAN), and capsule networks (CapsNet), and explored the use of deep learning in the identification of crop diseases and insect pests. In the review, they also compared the advantages and disadvantages of traditional machine learning and deep learning in the application of pest identification, and concluded that deep learning has stronger accuracy and generalization in pest identification. Zhong Linyi and others collected a total of 200 image samples of litchi anthracnose [6], litchi acid rot, litchi stink bug, litchi felt disease and other 10 types of litchi pests and diseases, using Inception v3 network and using transfer learning to train the last three layers of the network. The highest recognition accuracy rate of 96.30% was obtained. Zhang Shanwen and others collected 1200 images of cucumber diseased leaves, and designed an 11-layer leNet convolutional neural network [7], which achieved a higher recognition rate of 90.32% than traditional feature extraction methods. Barbedo and others used images of individual lesions and spots instead of images of whole leaves to identify 79 diseases of 14 plants [8], and obtained an average accuracy rate of 12% higher than that of the original image. Chen et al. first used an enhanced artificial neural network to segment the image, and then input the segmented image into the convolutional neural network, and obtained a recognition accuracy of 93.75% [9].

Deep neural networks have higher accuracy and generalization, and have broader application prospects in the identification of crop diseases and insect pests [10]. With

the support of transfer learning, the training of deep neural network models has been accelerated, and the average recognition accuracy has also been increased.

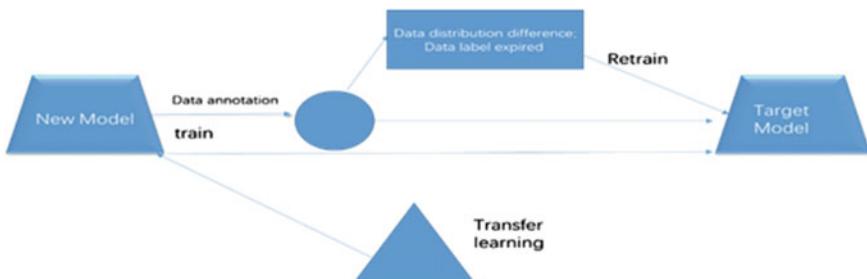
### 3 Introduction to Transfer Learning and Residual Networks

#### 3.1 Transfer Learning

We know that traditional machine learning requires a lot of labeling when training models, and labeling data often takes a lot of time. At the same time, in actual application scenarios, there are often problems of label data expiration and data distribution differences caused by the expiration of training data. Transfer learning refers to the transfer of pre-trained model parameters to the new model to help the training of the new model [11]. Therefore, this can not only greatly save the time spent on data labeling, but also improve the training accuracy and speed up the model optimization efficiency (Fig. 1).

There are currently three main ways of transfer learning:

- (1) Fine-tuning: Freeze part of the convolutional layer of the pre-trained model (usually most convolutional layers close to the input, because these layers retain a lot of underlying information) or even do not freeze any network layers, and train the remaining convolutional layers (usually the part close to the output) Convolutional layer) and fully connected layer [12].
- (2) Transfer Learning: Freeze all the convolutional layers of the pre-trained model, and only train your own customized fully connected layers.
- (3) Extract Feature Vector: First calculate the feature vector of the convolutional layer of the pre-training model for all training and test data, and then abandon the pre-training model, and only train your own customized simplified version of the fully connected network.



**Fig. 1** Transfer learning

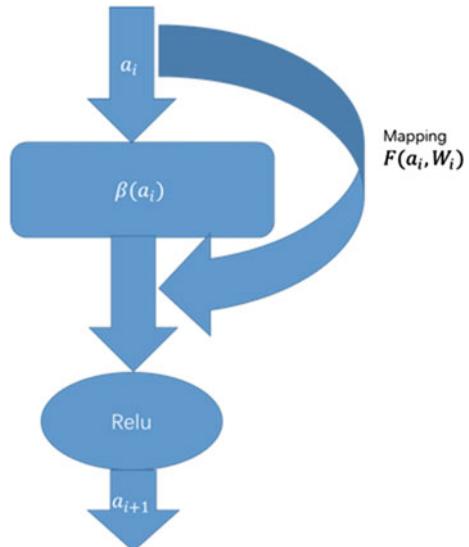
### 3.2 Residual Networks

The residual network is composed of residual blocks, one of which can be expressed mathematically as:

$$a_{i+1} = \beta(a_i) + F(a_i, W_i) \quad (1)$$

A residual block is divided into direct mapping and residual part, that is  $\beta(a_i)$  and  $F(a_i, W_i)$ , which  $\beta(a_i)$  in the expression is direct mapping and  $F(a_i, W_i)$  is the residual part. The residual network is proposed to solve the phenomenon of reduced accuracy as the number of network layers increases. In the construction of a convolutional neural network, the more the number of network layers, the loss will first decrease and then reach saturation. When the loss is in a saturated state, increasing the number of layers of the network will cause the loss to decrease. This phenomenon is called degradation. When it is in the degradation phenomenon, the low-level network has better effects than the high-level network. Therefore, the idea of connecting different network layers through direct mapping has become a key means to solve the degradation phenomenon, and the residual network has been on the stage of history (Fig. 2).

**Fig. 2** Residual block





**Fig. 3** Display of corn status

## 4 Data Processing and Model Building

### 4.1 Data Set

The data taken in the experiment is the corn growth state data set, which includes four corn states: health, spot disease, small spot disease and crust. The number of pictures corresponding to each state is 433, respectively, 354, 187, 432, a total of 1406. The picture is shown in Fig. 3.

### 4.2 Data Processing

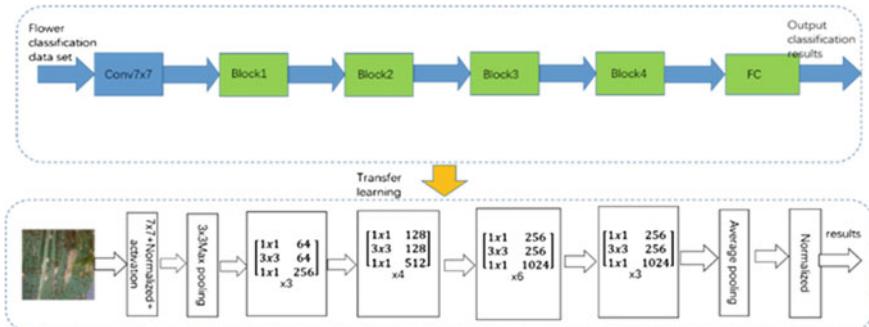
In order to improve the performance of the model, this paper first enhances the four types of data, including random resize cropping to  $256 * 256$  specifications, random rotation, random horizontal flipping, image center cropping to  $224 * 224$ , conversion

to Tensor and normalization. Then through the script file, the original four types of picture sets are divided into training set and validation set according to 9:1.

### 4.3 ResNet50 Model Construction Based on Transfer Learning

Since the residual network was proposed, it has achieved great influence in the field of image recognition. Typical residual networks include ResNet50, ResNet101, ResNet152, etc. This paper selects ResNet50 and then uses the classic flower classification model for transfer learning to build the recognition of corn diseases and insect pests Model.

The transfer learning model of classic flower classification is shown in Fig. 4. In the pre-training model of flower classification, the data will firstly go through a  $7 * 7$  convolution operation, and then go through four layers of blocks (residual blocks). Perform operations such as normalization, pooling and convolution on the data stream, and finally through the FC layer for average pooling and SoftMax output classification results. In the residual block, the original part of the data is directly transported to the next layer through the identity mapping by learning the residual, so as to avoid the loss of information and improve the performance of the network to a certain extent.



**Fig. 4** ResNet50 model based on transfer

## 5 Model Training and Result Analysis

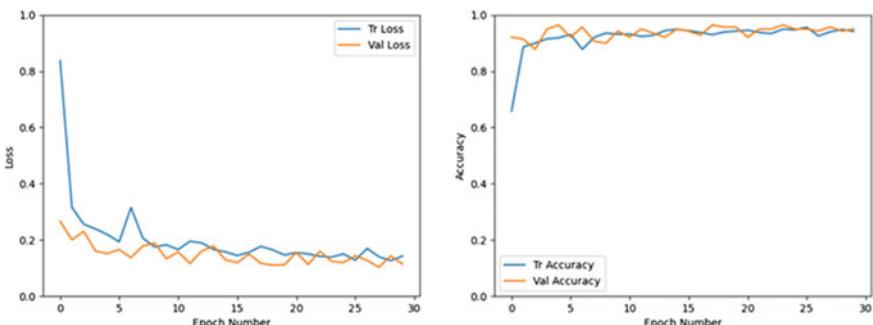
### 5.1 Model Training

This experiment uses the pycharm compiler under the pytorch framework to execute. At the same time, in order to make the experiment better, the GPU running model GTX1650 is used, and VGG16 and InceptionV3 are used for comparison experiments. When loading data, the DataLoader package is introduced. Using the DataLoader package not only makes it easier to load data, but also allows the data stream obtained from the data enhancement operation in the data preprocessing stage to be input into the model. In the introduction of transfer learning, in order to make it more suitable for this model, we replaced the last layer of ResNet50 with the input of the original last fully connected layer to a linear layer with 256 output units, and then connected the ReLU layer and Dropout Layer, then a  $256 \times 10$  linear layer, and the output is a 10-channel softmax layer. In the training process, we use the Adam optimizer, and the number of iterations is 30.

### 5.2 Experimental Results and Analysis

As shown in Fig. 5, when ResNet50 based on transfer learning recognizes the four-classification data set of corn diseases and insect pests, as the loss curve on the training set and the validation set (Fig. 5 left) declines and tends to be consistent, the training set The prediction accuracy rate of the model on the validation set has also reached the same level and no longer increases. At this time, the model has reached the fitting state and the final prediction accuracy is 96.42%.

As shown in Table 1, compare the accuracy rates obtained by VGG16, ResNet50 and ResNet50 based on transfer learning in the corn four-category pest data set. For the three models of VGG16 and ResNet50, the residual network ResNet50 has



**Fig. 5** Loss and accuracy curves of training set and validation set

**Table 1** Prediction accuracy rate of the four types of models on the corn pests and diseases data set

	ResNet50 based on transfer learning	VGG16	ResNet50
Training set/validation set (%)	96.42	92.4	94

the highest prediction accuracy on the training set and the validation set; and the ResNet50 model that uses transfer learning shows better results than ResNet50 that does not use transfer learning.

## 6 Conclusion

The recognition of corn pests and diseases through deep learning has the advantages of high efficiency, high accuracy and fast speed. This paper uses the ResNet50 model based on transfer learning to practice the corn pests and diseases data set, and the prediction accuracy rate reaches 96.42%. At the same time, comparing VGG16 and ResNet50 which does not use transfer learning, it is concluded that the residual network and transfer learning have the characteristics of high efficiency and high accuracy in the prediction process.

## References

1. Kecheng, B., Haijun,Y., Yonghua, L.: A review of the application of deep learning in the detection and recognition of agricultural pests and diseases **20**, 26–33 (2021)
2. Xuewei, J., Xingying, H., Xue, D., Xiaoping, W.: Crop pests and diseases identification method based on deep learning **51**, 182–183 (2020)
3. Shaopeng, J., Hongju, G., Xiao, H.: Research progress in image recognition technology of crop diseases and insect pests based on deep learning **50**, 313–317 (2019)
4. Dongfang, W., Jun, H.: Crop disease classification based on migration learning and residual network **37**, 199–207 (2021)
5. Lei, H.: Single image super-division reconstruction based on multi-level perceptual residual convolutional network **26**, 776–786 (2021)
6. Linyi, Z., Haifeng, L., Lizhong, D., Xiang, G., Jiayi, H., Zhiheng, S.: Research on image recognition of crop diseases and insect pests under computer vision **42**, 51–55 (2021)
7. Shanwen, Z.: Application of convolutional neural network in the recognition of cucumber leaf diseases **34**, 56–61 (2018)
8. Barbedo, J.G.A.: Plant disease identification from individual lesions and spots using deep learning **180**, 96–107 (2019)
9. Junde, C., Jinxiu, C., Defu, Z., et al.: A cognitive vision method for the detection of plant disease images **32**, 18 (2021)

10. Chunmei, D., Taochuan, Z.: Research on surface quality detection of preforms based on residual network and migration learning **143**, 138–139 (2021)
11. Bingyan, D., Zhishuo, Z.: Research on recyclable waste recognition and classification based on transfer. Learning **40**, 94–100 (2020)
12. Aoyu, L., Yunzhi, W., Xiaoning, Z., Guohua, F.: Maize disease recognition based on deep residual network **37**, 67–74 (2021)

# A Lightweight Image Super-Resolution Network Based on ESRGAN for Rapid Tomato Leaf Disease Classification



Lei Zha , Yangjing Shi , and Juan Wen 

**Abstract** The Crop disease identification is an important task in intelligent agriculture. Image resolutions have a large impact on the overall accuracy of classification performance. Some crop diseases are so similar that low-resolution images cannot capture their differences. To reduce the losses caused by crop diseases, it is vital to study crop image super-resolution reconstruction. Recently, with the rapid development of deep learning, various Single Image Super-Resolution (SISR) methods based on convolutional neural network (CNN) have achieved remarkable performance. However, the existing SR networks mainly have large parameter sizes, which require numerous training images and computing resources. In this paper, a lightweight image super-resolution model is constructed and applied to tomato leaf disease identification. We introduce the Shuffle Blocks with an attention mechanism to replace the Residual in Residual Dense Blocks (RRDBs) in the Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN), which is an effective SR model and won first place in the PIRM2018-SR Challenge. By the special structure we designed, our model can significantly reduce the parameter size of ESRGAN while almost maintaining its performance. Besides, we employ the tomato leaf images from the Plant Village dataset to train and test our model. Finally, we use the VGG16 to classify tomato leaf diseases based on the reconstructed images. The experiment results show that our model can effectively reduce the parameter size, computational complexity, and image reconstruction time compared to other chosen SR networks. Furthermore, the accuracy of SR images generated by our model is closer to ESRGAN and higher than other state-of-the-art methods.

**Keywords** Tomato disease classification · SISR · Plant village · ESRGAN · VGG16

---

L. Zha · Y. Shi · J. Wen ()

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

e-mail: [wenjuan@cau.edu.cn](mailto:wenjuan@cau.edu.cn)

## 1 Introduction

Crop diseases are a major threat to food production and security [1, 2]. As one of the most important agricultural countries in the world, China has been facing significant issues in crop disease prevention and control. The rapid and accurate identification of crop diseases is the first and the most critical step of crop disease prevention and control. However, rapid and accurate identification remains challenging due to the lack of the necessary assistance. One of the factors affecting the identification accuracy is image resolution. Due to the limitations of the technology, equipment, and environment, the crop leaf images collected by certain devices are not always of good quality and not conducive to crop disease identification tasks. Therefore, it is a meaningful and challenging task to improve the quality and resolution of the crop images.

So far, the traditional image Super-Resolution (SR) reconstruction methods [3–5] have made a lot of progress. However, these traditional methods still have many defects, such as complex and inefficient image preprocessing processes. With the rapid development of deep learning and convolutional neural networks (CNN), many deep-learning-based super-resolution methods have achieved considerable performance and gradually replaced traditional SR methods.

In 2014, Dong et al. [6] first brought CNN into the SISR task and presented Image Super-resolution Convolutional Neural Network, termed SRCNN. Later, they proposed the Accelerating Super-Resolution Convolutional Neural Network (FSRCNN) [7], which improved the SRCNN in terms of reconstruction speed and image quality. Inspired by the residual network ResNet [8] proposed by He et al. in 2015, Kim et al. [9] deepened the number of network layers to 20 and came up with the VDSR model in 2016. Although the methods mentioned above achieved better results, they were not able to obtain finer texture details. Until 2017, SRGAN [10] proposed by Ledig et al. applied the generative adversarial network to the SISR issue for the first time, which achieved a good visual effect. Subsequently, Wang et al. [11] particularly removed the batch normalization layer in SRGAN and presented Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN), which won the championship in the PIRM2018-SR Challenge.

From SRCNN with three convolutional layers to ESRGAN with 350 layers, the depth and entire performance of the network have been markedly increased. However, even though the deep networks improved the quality of the SR images, sometimes it may not be suitable for the actual scene due to the low efficiency. To achieve rapid detection, it makes sense to use a lightweight model that is practical for the application. Considering the difficulty and efficiency of the model training and the applicability of the agricultural scenarios, we design a lightweight Super-resolution network based on the ESRGAN, namely LESRGAN, by introducing the Shuffle Blocks [12] with an attention mechanism to substitute the Residual in Residual Dense Blocks (RRDBs) in ESRGAN. With the help of the leaf image reconstructed by the lightweight model, the accurate and fast detection of leaf disease can be realized.

The contributions of this study can be concluded as below:

- We presented an efficient block called Shuffle Squeeze and Excitation Block (SSEB), which is used to replace the Residual in Residual Dense Block (RRDB) in ESRGAN. By combining the SE module and the ShuffleNetV2 unit, SSEB can learn the information effectively in the lightweight network.
- We employ LESRGAN to reconstruct low-resolution tomato leaf images and apply the VGG16 classifier to identify tomato leaf diseases. The experimental results show that LESRGAN effectively reduces the model scale. Furthermore, the image reconstructed by LESRGAN is beneficial for fast and accurate tomato disease identification.

The rest of this paper is as organized as below. In Sect. 2, we present the related works of our paper. In Sect. 3, we introduce LESRGAN in detail. In Sects. 4 and 5, experimental results and analysis are provided. Finally, we make a conclusion in Sect. 6.

## 2 Related Works

### 2.1 Single Image Super-Resolution Methods Based on CNN

Since the precursory work SRCNN [6] was proposed, plentiful methods based on deep learning have achieved remarkable improvement in SISR. Dong et al. [7] proposed FSRCNN to speed up and improve the SRCNN. Kim et al. [8] presented a very deep convolutional network named VDSR, verifying that adding the network depth is beneficial to enhancing the SISR performance to some extent. A year later, Enhanced Deep Residual Networks for Single Image Super-Resolution (EDSR) [13] was proposed by Lim et al. It was further proved that increasing the depth of SR network will improve the SR effect. Particularly, in 2014, Goodfellow et al. [14] first proposed the Generative Adversarial Network (GAN), which is consisted of the generator and discriminator. Inspired by GAN, Ledig et al. [10] first employed the GAN in SISR task and presented SRGAN, which used the perceptual loss to reduce the difference between human visual perception and Super-resolution images. Afterward, Wang et al. [11] improved the SRGAN by bringing the Residual-in-Residual Dense Block (RRDB) to the network of the generator and proposed Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN), which used the discriminator to predict relative realness rather than the common absolute difference in SISR. As a result, ESRGAN achieved a remarkable result and won the first in the Competition named PIRM2018-SR Challenge.

## 2.2 ShuffleNet

In order to apply the neural networks on mobile devices, Zhang et al. [15] designed the ShuffleNet V1 model, which cut the calculation consumption enormously. In ShuffleNet V1, channel shuffle operation was proposed to make up the defect of the pointwise group convolution. For the purpose of building a more lightweight model, Ma et al. [12] then proposed the ShuffleNet V2. They pointed out the bottleneck units and pointwise group operation would add the memory access cost. Besides, a large number of groups would lower the parallelism. To address these problems, Ma et al. [12] proposed the Channel Splitting to substitute the operation of group, which divided the input information into two embranchments. Specifically, they used the channel shuffle operation to promote the full exchange of information.

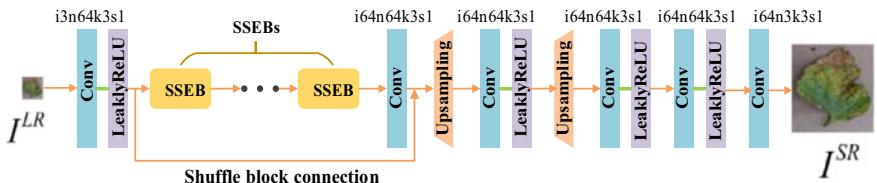
## 3 Our Method

### 3.1 SISR Model

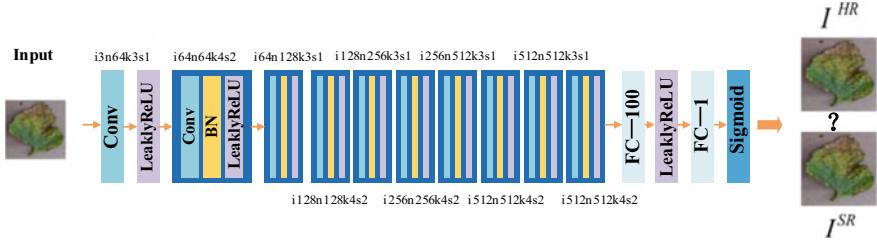
We use the basic framework of the ESRGAN. The whole network consists of a generator and a discriminator.

**Generator network.** We replace the Shuffle Squeeze and Excitation Blocks (SSEBs) with RRDBs, which are the basic block of the ESRGAN. In Fig. 1, the  $i$ ,  $n$ ,  $k$ , and  $s$  in the upper middle of each layer represent the number of input channels, the number of the output channels, the convolution kernel size, and the stride, respectively. The  $I^{LR}$  is the input of the model, which is sent to a  $3 \times 3$  convolution. The input and output channels are 3 and 64, respectively. The LeakyReLU [16] is used as the activation function. Subsequently, the feature information is feed into the SSEBs, which is the main feature extraction module. The structure of SSEB will be described in detail in Sect. 3.1.3. After SSEBs, the SR images are obtained through 5 convolutions and 2 upsampling operations.

**Discriminator network.** Keep the same as ESRGAN, we use the discriminator module in RaGAN [17] as our discriminator network, which is used to determine



**Fig. 1** The architecture of the generator

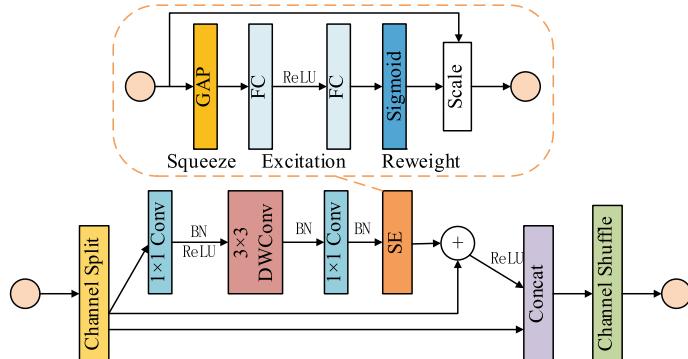


**Fig. 2** The architecture of the discriminator

which image is more real. We denote it as  $D_{R_a}$ . As is shown in Fig. 2, the feature extraction part consists of 10 convolution layers. The convolution layers with convolution kernels of  $3 \times 3$ ,  $4 \times 4$  appear in an alternate manner. BN [18] represents the Batch Normalization layer, FC is the fully connected layer. The specific parameters of each convolution layer have been indicated in Fig. 2.

**Shuffle Squeeze and Excitation Block (SSEB).** As the most important structure in the ESRGAN, Residual in Residual Dense Block (RRDB) uses a large number of dense connection blocks, which may transmit a lot of redundant feature information and increase the network complexity. To address these problems, we introduce the design concepts of ShuffleNet V2 into the ESRGAN model framework. Besides, we employ the channel attention mechanism of the SE network [19] to combine the residual learning. In this way, we propose the SSEB module, which is used to displace the RRDB.

The structure of SSEB is illustrated in Fig. 3, DWConv stands for deep convolution layer, FC represents the fully connected layer, and global average pooling (GAP) stands for global average pooling layer. At the beginning of each SSEB module, the operation of the channel split is performed. Then, the feature map is divided into two branches where the lower part of the branches uses the residual structure to



**Fig. 3** The architecture of the SSEB

directly connect the Contact Layer. The upper branch is further divided into convolution branch and residual connection. These two branches are finally added element by element, and it is connected to the bottom residual branch after the ReLu [20] activation function. The upper part of the convolution branch contains 3 convolution layers and a SE module. Through the operation of the contact, the number of input channels is equal to the output one so that the amount of memory access can be minimized. Finally, the channel shuffle manipulation is employed to further enhance the exchange of information.

*Squeeze and Excitation module.* The SE module is mainly divided into three parts: Squeeze, Excitation, and Reweight. As is illustrated in Fig. 3, the GAP is introduced in the squeeze section to add global spatial information based on channel dimension. Additionally, in the excitation section, two fully connected layers are employed to collect the information that comes from the previous part. The excitation part is to capture the information dependency between channels. Finally, the Sigmoid function is used to obtain the normalized weights between 0 and 1, which will be allocated to each channel of the input feature maps.

The number of SSEB modules needs to be determined experimentally, which will be discussed in detail in Sect. 5.1.

### 3.2 Loss Functions

**Loss function for Discriminative network.** As is illustrated in Fig. 2, the discriminator  $\mathbf{D}_{Ra}$  can get two outputs, which are  $\mathbf{D}_{real}$  and  $\mathbf{D}_{fake}$ . The equations of them can be expressed as:

$$\mathbf{D}_{real} = \mathbf{C}(\mathbf{I}^{HR}) - \mathbf{E}(\mathbf{C}(\mathbf{I}^{SR})) \quad (1)$$

$$\mathbf{D}_{fake} = \mathbf{C}(\mathbf{I}^{SR}) - \mathbf{E}(\mathbf{C}(\mathbf{I}^{HR})) \quad (2)$$

where  $\mathbf{D}_{real}$  represents the average probability that the result of  $\mathbf{D}_{Ra}$  is the Ground Truth (HR image).  $\mathbf{D}_{fake}$  means the average probability that the result of  $\mathbf{D}_{Ra}$  is the SR image.  $\mathbf{C}(\cdot)$  is the function of the discriminator.  $\mathbf{E}(\cdot)$  represents the function of getting the average from the min-batch data.

Thus, we mark the loss of the discriminator as  $\mathbf{L}_D^{Ra}$ , which is divided into two parts: real loss  $\mathbf{L}_{D_{real}}^{Ra}$  and fake loss  $\mathbf{L}_{D_{fake}}^{Ra}$ . The aim of real loss is to make the real images more realistic than the fake ones, and the fake loss is employed to make the fake images less realistic than the real ones. The equation of the  $\mathbf{L}_D^{Ra}$  can be expressed as follows:

$$\mathbf{L}_D^{Ra} = \mathbf{L}_{D_{real}}^{Ra} + \mathbf{L}_{D_{fake}}^{Ra} \quad (3)$$

$$L_{D_{real}^{Ra}} = -E_{I^{HR}} [\log(D_{Ra}(I^{HR}, I^{SR}))] \quad (4)$$

$$L_{D_{fake}^{Ra}} = -E_{I^{SR}} [\log(1 - D_{Ra}(I^{SR}, I^{HR}))] \quad (5)$$

where  $D_{Ra}(I^{HR}, I^{SR}) = \sigma(C(I^{HR} - E_{I^{SR}}(C(I^{SR}))))$ ,  $D_{Ra}(I^{SR}, I^{HR}) = \sigma(C(I^{SR} - E_{I^{HR}}(C(I^{HR}))))$ .  $\sigma$  is Sigmoid activation function.

**Loss function for the generative network.** We use the sum of the perceptual loss function [21], content loss, and adversarial loss as our loss function for the generator network.

- (1) The perceptual loss is denoted as  $L_{perceptual}$  and it is defined on the feature graph before the ReLu activation function of the pre-trained VGG19 model. The formula is expressed as:

$$L_{perceptual} = \frac{1}{W_{5,4}H_{5,4}} \sum_{x=1}^{W_{5,4}} \sum_{y=1}^{H_{5,4}} (\Phi_{5,4}(I^{HR})_{x,y} - \Phi_{5,4}(G(I^{LR}))_{x,y})^2 \quad (6)$$

where  $\Phi_{5,4}$  represents the pre-trained VGG19 network, using the feature maps after the 4th convolution layer and before the 5th largest pooling layer, which stand for the high-level semantic features and similarities.  $W_{5,4}$ ,  $H_{5,4}$  are the width and height of the feature map, respectively.  $G(\cdot)$  is the function of our generator.

- (2) Then the content loss is used to evaluate the absolute difference between the SR and HR. We denote the content loss as follows:

$$L_1 = E_{I^{SR}} I^{SR} - I^{HR} \quad (7)$$

- (3) The aim of adversarial loss is to encourage the network to support the solutions, which exist in the manifold of the images. According to the  $L_D^{Ra}$  of the discriminator, the equation of the adversarial loss  $L_G^{Ra}$  used in the generator can be expressed as:

$$L_G^{Ra} = -E_{I^{HR}} [\log(1 - D_{Ra}(I^{HR}, I^{SR}))] - E_{I^{SR}} [\log(D_{Ra}(I^{SR}, I^{HR}))] \quad (8)$$

Thus, the total loss of the generative network can be formulated as an equation:

$$L_G = L_{perceptual} + \alpha L_G^{Ra} + \beta L_1 \quad (9)$$

where  $\alpha$ ,  $\beta$  are two coefficients used to balance between various loss terms.

## 4 Experiments

### 4.1 Experiments Setup

Our experiments mainly use PyTorch1.1.0 and Tensorflow1.3.1 with python3.5 as deep learning frameworks on a server running Ubuntu16.04. The running memory of the server is 23.5G, and the disk size is 1.7 T. And we utilize one NVIDIA GeForce GTX 1070 GPU to accelerate the calculation. In addition, MATLAB 2018a is also used on the host of the Windows 10 system to process some images and calculate the evaluation index PI [22].

### 4.2 Datasets

The dataset is a total of 18,160 tomato images in Plant Village [23], with 9 kinds of leaf diseases plus healthy one, the name and quantities of each category are shown in Table 1. These image sizes have been uniformly processed to  $256 \times 256$ . Due to limited computing resources, we cropped each image to a size of  $128 \times 128$ . Furthermore, we randomly divide the whole dataset into training set, validation set, and test set according to 8:1:1. As a result, our dataset consists of 14,536 images, 1812 validation images, and 1812 testing images. We compare the performances on our test dataset.

We use the bicubic kernel function by downsampling the original high-resolution images to obtain the low counterparts. All experiments are conducted on  $\times 4$  scale factor.

**Table 1** The name, label, and number of images of each kind of tomato leaf in the Plant Village dataset

No.	Classification name	Label	Number
0	Xanthomonas campestris pv. Vesicatoria	Bacterial	2027
1	Alternaria solani	Early blight	1000
2	Phytophthora Infestans	Late blight	1909
3	Fulvia fulva	Mold leaf	952
4	Septoria lycopersici	Septoria	1771
5	Tetranychus urticae	Spider mites	1676
6	Corynespora cassiicola	Target spot	1404
7	Tomato yellow leaf curl virus	Curl virus	5357
8	Tomato mosaic virus	Mosaic virus	373
9	Healthy	Healthy	1591

### 4.3 Training Details

To be specific, we train the generator using the loss function in Eq. (9) with  $\alpha = 5 \times 10^{-3}$ ,  $\beta = 1 \times 10^{-1}$ . And we train our generator and discriminator network with Adam [24] optimizer by setting  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ . The initialized learning rates of the discriminator and generator are both  $1 \times 10^{-3}$ , which will be reduced by half every 50 k iterations. The maximum number of iterations is set to 400 k. Due to limited computing resources, the batch size is set to 16.

## 5 Experiments Results and Analysis

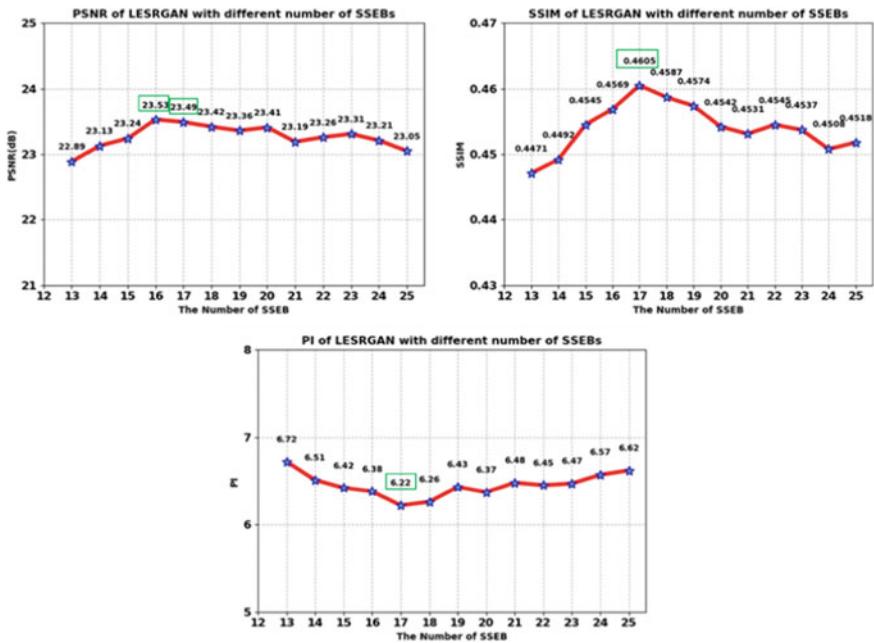
### 5.1 Selection of the Number of SSEBs

The key variable of LESRGAN is the number of SSEB modules. Therefore, we explore the impact of the number of SSEBs on the three evaluation indicators of the SR models (PSNR, SSIM [25], and PI). We set the number of SSEBs from 13 to 25 and train our models on the tomato leaf dataset. Then the average PSNR and SSIM are calculated on all the test sets while the calculation of the PI is still on the subset of the tomato leaf test set. Each type of tomato leaf disease contains 10 images, so the total of 100 images are used to compute the PI. As shown in Fig. 4, as the number of SSEBs increases, the average of the PSNR and SSIM also gradually improve, and PI keeps decreasing. However, after the number reaches a certain value, the trend of three curves will go in the opposite direction. This is because when the number of SSEBs is small, the advantages of the model cannot be reflected. But when the number is too large, it will lead to a huge solution space for the model, which means that the model is difficult to converge to the best state. In particular, when the number reaches 17, the SSIM reaches the highest value of 0.4605, and the PI is as low as 6.22. As a result, we set the number of SSEBs as 17.

### 5.2 The Comparison with Different Methods

In our SR experiments, six different methods are used for comparison (shown in Table 2). Among them, the ESRGAN adopts a two-step transfer learning training strategy. The specific details have been implemented in our previous work [26].

Although ESRGAN has a good performance in both objective and subjective evaluation of reconstructed image quality, it has a deep network structure with a large number of parameters, leading to high computational complexity. By contrast, our LESRGAN realizes lightweight while maintaining performance to a certain extent. As you can see in Table 2, the number of parameters and calculations of LESRGAN have been reduced significantly compared to those of ESRGAN. More importantly,



**Fig. 4** The influence of the number of SSEBs on the evaluation indicators PSNR, SSIM, and PI

**Table 2** PSNR, SSIM, PI, Parameters, and FLOPs for scale  $\times 4$

Method	PSNR	SSIM	PI	Parameters	FLOPs
Bilinear	24.45	0.4776	7.23	—	—
SRCNN	25.42	0.5126	7.16	57 K	52.7 G
FSRCNN	25.58	0.5415	7.20	12 K	6.0 G
SRGAN	23.44	0.4495	6.52	1.55 M	169.2 G
ESRGAN	23.70	0.4678	6.10	16.7 M	1177 G
LESRGAN (our)	23.49	0.4605	6.22	0.21 M	94.48 G

the value of PI of LESRGAN is the second-lowest among all six methods. Additionally, we calculate the time for reconstructing the input  $64 \times 64$  low-resolution image into a  $256 \times 256$  image by ESRGAN and LESRGAN. The test results show that ESRGAN processes each image in an average of 49.86 ms, while LESRGAN needs 16.16 ms, which reduced the time by two-thirds.

### 5.3 SR Visual Effect

We show the qualitative comparisons over the tomato leaf datasets, including 0379 from the late\_blight dataset, 0452 from the Septoria dataset, and 0268 from the early\_blight dataset. The visual comparisons for  $\times 4$  SR images are utilized to compare the reconstruction performance. As you can see in Fig. 5, our LESRGAN performs close to the ESRGAN and better than other methods. In addition, the result of our model is more natural compared to the Bilinear, SRCNN, FSRCNN, and SRGAN.

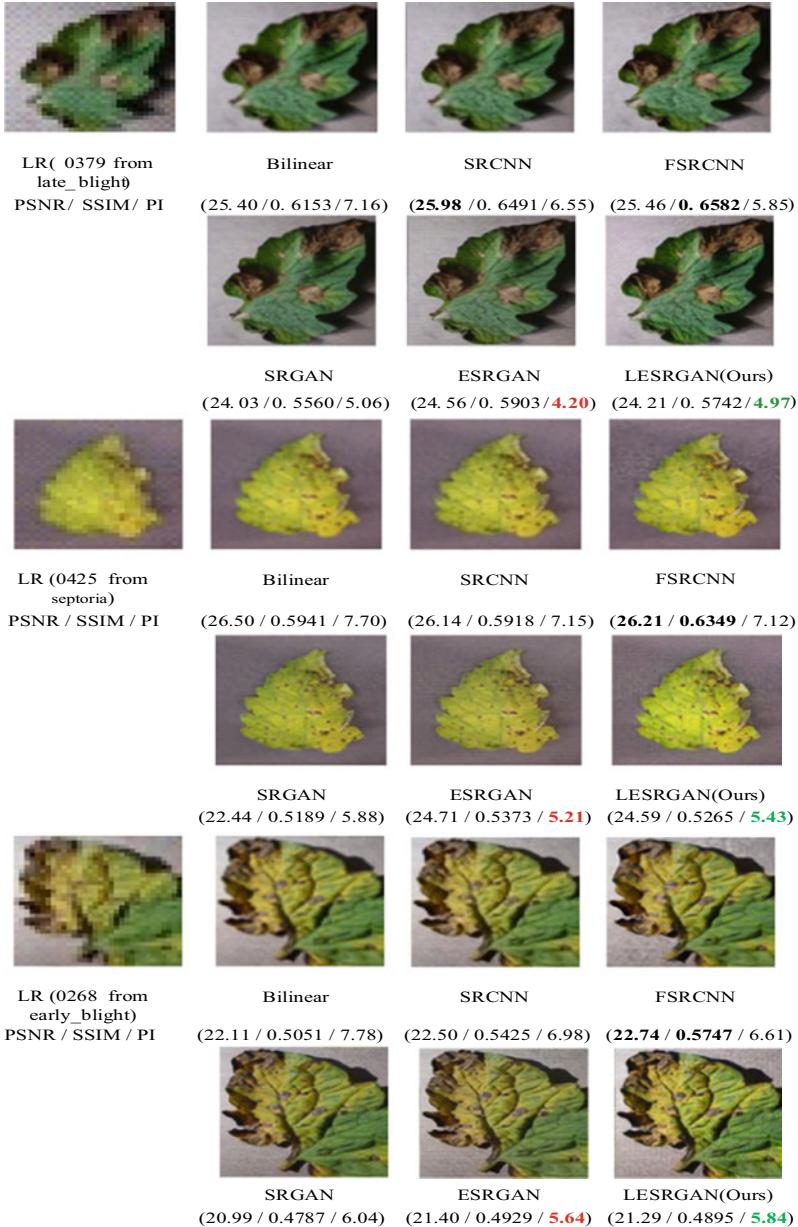
### 5.4 Tomato Leaf Disease Classification

We use the standard VGG16 [27] as the classification model for tomato leaf disease super-resolution images. The VGG16 model is pretrained on ImageNet [28], then fine-tuned by the SGD optimizer with a stable learning rate of  $5 \times 10^{-4}$ . Our model ultimately converged after 10 k iterations. The accuracy comparison of different model classifications is shown in Table 3.

The results show that the classification accuracies obtained by the reconstructed SR images via ESRGAN and LESRGAN are higher than those of other models, which are closest to the classification accuracy of HR. However, the difference is still about 4–5%. The classification accuracies of SR images generated by ESRGAN are the highest, reaching 85.38 and 90.78%. The classification accuracies of LESRGAN are 0.94 and 0.49%, slightly lower than the former, respectively.

## 6 Conclusion

In this paper, to better identify tomato leaf diseases, we presented a novel lightweight Super-Resolution network based on ESRGAN. We combine the ShuffleNet V2 unit and the SE module, including channel attention mechanism and residual structure, to improve ESRGAN. Our method achieved lightweight under the premise of ensuring the quality of the SR images. The experiments show that our LESRGAN is slightly lower than ESRGAN in the three image evaluation indicators of PSNR, SSIM, and PI, but it is better in terms of the number of model parameters, calculations, and the reconstruction time. Ultimately, the implementation shows that the accuracy of crop disease classification with images generated by LESRGAN is close to those by ESRGAN but better than those by other methods.



**Fig. 5** Visual contrast for  $\times 4$  SR on tomato leaf dataset (the bold shows the best value of PSNR or SSIM. The red and the green represent the best and the second best of PI, respectively)

**Table 3** Comparison of the classification results of LR, SR, and HR on tomato leaf dataset

	Accuracy (%) (LR image size is 32×32, other image sizes are 128×128)	Accuracy (%) (LR image size is 64×64, other image sizes are 256×256)
LR	52.65	69.81
Bilinear	71.14	82.40
SRCNN	80.13	86.09
FSRCNN	80.85	87.20
SRGAN	82.78	88.41
ESRGAN	<b>85.38</b>	<b>90.78</b>
LESRGAN	<b>84.44</b>	<b>90.29</b>
HR	89.96	95.14

## References

- Celia, A.H., Zo, L.R., Nalini, S.R., Radhika, D., Hery, R., Rivo, H.R., Haingo, R., James, L.M.: Extreme vulnerability of smallholder farmers to agricultural risks and climate change in Madagascar. *Philos. Trans. R. Soc. Lond. Ser. A* **369**, 20130089 (2014)
- Amos, P.K.T., Maria, V.M., Colette, L.H.: Threat to future global food security from climate change and ozone air pollution. *Nat. Clim. Chang.* **4**, 817–821 (2014)
- Jianchao, Y., John, W., Thomas, S.H., Yi, M.: Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **19**(11), 2861–2873 (2010)
- Jianchao, Y., Zhaowen, W., Zhe, L., et al.: Coupled dictionary training for image super-resolution. *IEEE Trans. Image Process.* **21**(8), 3467–3478 (2012)
- Radu, T., Rasmus, R., Luc, van G.: Seven ways to improve example-based single image super resolution. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1865–1873 (2016)
- Chao, D., Chen-Change, L., Kaiming, H., Xiaou, T.: Learning a deep convolutional network for image super-resolution. In: *ECCV* (2014)
- Chao, D., Chen-Change, L., Xiaou, T.: Accelerating the super-resolution convolutional neural network. In: *ECCV* (2016)
- Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. In: *CVPR* (2016)
- Jiwon, K., Jung-Kwon, L., Kyoung-Mu, L.: Accurate image super-resolution using very deep convolutional networks. In: *CVPR* (2016)
- Christian, L., Lucas, T., Ferenc, H., Jose, C., Andrew, C., Alejandro, A., Andrew, A., Alykhan, T., Johannes, T., Zehan, W., Wenzhe, S.: Photo-realistic single image super-resolution using a generative adversarial network. In: *CVPR* (2017)
- Xintao, W., Ke, Y., Shixiang, W., Jinjin, G., Yihao, L., Chao, D., Yu, Q., Chen-Change, L.: ESRGAN: enhanced super-resolution generative adversarial networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 63–79. Munich, Germany (2018)
- Ma, N., Zhang, X., Zheng, H., Sun, J.: ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: *European Conference on Computer Vision* (2018)
- Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: *CVPRW* (2017)
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *NIPS* (2014)

15. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. *Comput. Vision Pattern Recognit.* (2018)
16. Kaiming, H., Xiangyu, Z., Shaoqing, R., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034 (2015)
17. Alexia, J.: The relativistic discriminator: a key element missing from standard GAN. In: *International Conference on Learning Representations* (2019)
18. Sergey, I., Christian, S.: Batch normalization: accelerating deep network training by reducing internal covariate shift. [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
19. Jie, H., Li, S., Samuel, A., Gang, S., Enhua, W.: Squeeze-and-excitation networks. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
20. Vinod, N., Geoffrey, E.H.: Rectified linear units improve restricted Boltzmann machines, In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814. Haifa, Israel, Omnipress (2010)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *ECCV* (2016)
22. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: 2018 PIRM challenge on perceptual image super-resolution. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany (2018)
23. David, P. H., Marcel, S.: An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *arXiv* 2015, [arXiv:1511.08060v2](https://arxiv.org/abs/1511.08060v2) (2015)
24. Diederik, P.K., Jimmy, L.B.: Adam: a method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
25. Zhou, W., Alan, C.B., Hamid, R.S., Eero, P.S.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process* **13**, 600–612 (2004)
26. Juan, W., Yangjing, S., Xiaoshi, Z., Yiming, X.: Crop disease classification on inadequate low-resolution target images. *Sensors* **20**(16), pp. 4601 (2020).
27. Karen, S., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
28. Jie, D., Wei, D., Richard, S., Lijia, L., Kai, L., Feifei, L.: ImageNet: a large-scale hierarchical image database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, Miami, FL, USA (2009)

# Fuzzy Image Processing Based on Deep Learning: A Survey



Shoucun Chen , Jing Zhang , and Tianchi Zhang

**Abstract** Fuzzy images exist in all imaging processes, including computer vision, photography and medical imaging. Many methods based on deep learning can deblur the image and have a good output. Medical image is a typical processing scene of Fuzzy image. Various methods based on Deep Learning continue to emerge, which provide a good solution to the difficult problems in the field of medical image processing. For example, automatic data tagging is expected to solve the problems of great diversity and high labor costs in manual labeling of medical images. The automation of network architecture design based on search strategy has been able to design a network comparable to that designed by researchers manually. Federated Learning can realize a distributed training neural network model without sharing private data, which is expected to solve the problem of data privacy in the field of medical image processing. In this paper, firstly, the frontier methods of fuzzy image processing based on Deep Learning are introduced. Secondly, we summarize the research of automatic data label and automatic network architecture design. Finally, the application of Federated Learning in the field of medical images is summarized.

**Keywords** Fuzzy image · Deep learning · Federated learning

## 1 Introduction

As image blurring exists widely in daily life, the problem of image deblurring has been concerned and studied in the last century. According to the properties of fuzzy kernel, image blurring can be divided into blind image deconvolution and non-blind

---

S. Chen · J. Zhang

School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

T. Zhang ()

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China

e-mail: [zhangtianchi@cqjtu.edu.cn](mailto:zhangtianchi@cqjtu.edu.cn)

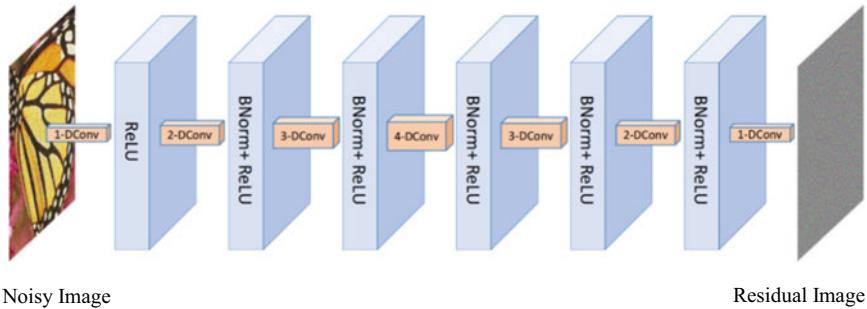
image deconvolution. BID recovers the image when the blur kernel is unknown. In this case, there is no information except the captured image. NBID is to restore a clear original image when the blur kernel is known. Most of the methods based on deep learning can propose an end-to-end image deblurring model and output a clear original image. In order to overcome the bottleneck of data labeling in the field of medical image processing. Active learning is a semi-supervised learning method that can reduce the cost of data tagging as much as possible. Generating an adversarial network is a generation model, which can generate real output similar to training data. Self-supervised learning can find the relationship between samples by mining the inherent features of the data, so as to make efficient use of unlabeled data. Federal Learning is expected to effectively aggregate the local knowledge acquired by institutions from private data, so as to further improve the accuracy, robustness and generalization of the depth model. These key research are very important to break through the bottleneck in medical image processing.

## 2 Fuzzy Image Processing Method Based on Deep Learning

For the problem of non-uniform blur of images, the current solutions are divided into traditional optimization methods and deep learning methods. However, for the blurring problem of dynamic scene, there is only blurring in the local area of the image, and some studies have proposed effective solutions based on deep learning.

The problem existing in the current image deblurring algorithm: it is difficult to obtain the measured clear image and blurred image pairs, which is used to train the image deblurring network. For the problem of image deblurring in dynamic scenes, it is difficult to obtain the blur kernel of local images. Getting rid of blurring the problem requires a greater sense of wildness. Nah et al. [1] proposed a method of image synthesis of measured dynamic scenes, and disclosed that deblurring data sets Gopro Large, Gopro Large data sets have become one of the commonly used data sets based on deep learning. The convolution neural network is used to restore the clear image directly from the degraded image, and according to the traditional image deblurring problem, the multi-scale restoration strategy is integrated into the network. Kupyn et al. [2] proposed to apply GAN to image deblurring and realized an end-to-end image deblurring based on deep learning. The generator consists of two convolution networks with a step size of 1ap2, nine ResBlock and two deconvolution networks. Each ResBlock includes a convolution layer, an instance normalization layer, and a ReLU activation layer. Kai et al. [3] proposed to use dilated filter to enlarge receptive field, using batch normalization and residual learning to accelerate training, using training samples with a small size to help avoid boundary artifacts, learning specific denoiser model with small interval noise levels. As shown in Fig. 1.

Dynamic scene deblurring is a challenging task in low-level vision. Different from the parameter independent or parameter sharing mode in the existing methods, they proposes a generalized and effective selective sharing mechanism to constrain the



**Fig. 1** The architecture of the proposed denoiser network

deblurring network [4]. In each scale subnetwork, a nested skip connection architecture is proposed to replace the residual module/convolution stack module. In addition, they build a larger deblurring dataset. Finally, the SOTA performance of the proposed nested skip join is verified by sufficient experiments. Some scholars have proposed that cyclic neural network can be used to solve the problem of motion blur in images. Zhang et al. [5] use the model of three convolution neural networks and one cyclic neural network to realize blind image de-blurring in this method, three convolution neural networks are used to extract feature images, train the weights needed to learn deconvolution and restore clear images according to the feature images obtained by deconvolution, while cyclic neural networks are used to perform deconvolution operations. Wang et al. [6] proposed a new image deblurring method, which can restore the naturally blurred image directly in the form of primary convolution filtering. They proposed a blind method to estimate the PFS statistics of two Gaussian and Laplace models, which are common in group multi-image transmission, and designed complete experiments to test and verify the effectiveness of the method, using 2054 natural blurred images, six imaging applications and seven advanced deconvolution methods.

### 3 Automatic Data Labeling Method in Deep Learning

Supervised learning can not avoid the dependence on labeled data, so automatic data generation is also a way to reduce labor costs. There are many ways to synthesize data, including manual design rules, using GAN network generation and so on.

### ***3.1 Generative Adversarial Networks Method for Data Annotation***

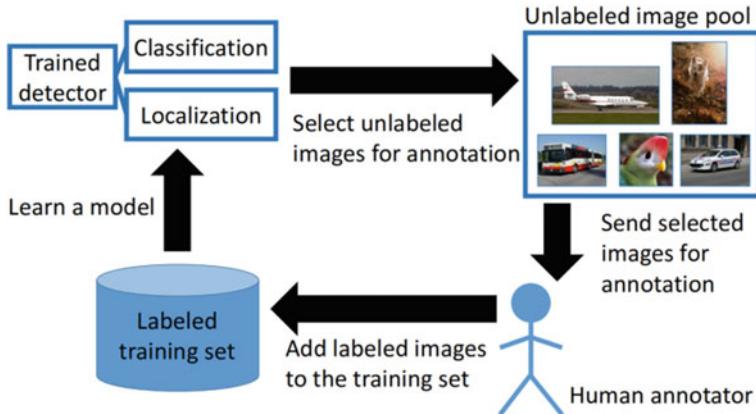
Among the methods of data synthesis, using GAN to synthesize domain adaptation is a research direction. The main concerns include the similarity between appearance and geometry of source domain and target domain. They consider two kinds of similarity at the same time to do generative confrontation [7]. Jaderberg et al. [8] proposed a synthetic data method based on manual design rules for text recognition tasks. The synthesized image sample is composed of foreground image layer, background image layer and shadow layer. The synthesis step is divided into six steps: Randomly select the font and present the text into the foreground layer, generate edge shadows from the text in the foreground layer, fill three layers, randomly distort the foreground and shadow, mix the image with the real scene image, add Gaussian noise and so on.

### ***3.2 Active Learning Method for Data Annotation***

Different samples help to improve the existing models differently, just like the human learning process, it is difficult for people who only learn primary school knowledge to break through the bottleneck of junior high school knowledge. In the classification problem, Zongwei et al. [9] proposed the active learning process in the classification task, which is measured according to the difference and uncertainty of the patch prediction of the input image generated by the model. Kao et al. [10] emphasized that the confidence of the detection box in the detection task only represents the classification confidence and does not have the location confidence, so it is proposed to supplement the location confidence to evaluate the advantages and disadvantages of the detection box. As shown in Fig. 2. Yoo et al. [11] point out that most of the existing active learning models are task-specific, so the sample selection strategy of task-agnostic is proposed, and the experiments are verified in classification, detection and other tasks.

### ***3.3 Self-supervised Learning Method for Data Annotation***

Self-supervised learning is a kind of unsupervised learning, which has recently become a research hotspot in academic circles. It uses the structure or characteristics of untagged data to artificially construct tags to supervise web-based learning. Usually, the self-supervised learning model is not directly applied to the target task, but as a pre-training model for downstream tasks. He et al. [12] proposed a new development of self-supervised learning. The effect of the unsupervised model obtained by using the method of this paper as a pre-training model after many downstream



**Fig. 2** A round of active learning for object detection

tasks fine-tune is better than that of using the supervised learning pre-training model fine-tune.

## 4 Automated Design of Network Architecture

Although the neural network has automated the troublesome feature extraction, the network structure still needs to be designed manually to a large extent. Every year, a lot of research work is done to propose a variety of new and better network substructures, so a natural demand is whether the work can be done by the machine.

The process of network architecture search is to define the search space first, then find out the candidate network structures through the search strategy, evaluate them, and conduct the next round of search according to the feedback.

### 4.1 Search Strategy Based on Reinforcement Learning

MIT researchers proposed MetaQNN [13], which models network architecture search as a Markov decision process and uses RL methods to generate CNN architectures. For each layer of CNN, you will learn to select the type of layer and the corresponding parameters. After the network structure is generated, the evaluation accuracy obtained after training is rewarded. Google researchers use the RNN network as the controller to sample and generate strings that describe the network structure, which is used to train and get the accuracy of the evaluation, and then use the reinforce algorithm to learn the parameters of the controller, so that it can produce a more accurate network structure. It uses 800 GPU, and finally beats the manual design model with similar

network architecture on the CIFAR-10 data set, reaches a new SOTA level on the PTB data set, and finds a better structure than the widely used LSTM [14].

## 4.2 Search Strategy Based on Evolutionary Algorithm

Evolutionary algorithms are introduced to solve the NAS problem, and it has been proved that high accuracy can be achieved from a simple initial condition on CIFAR-10 and CIFAR-100 data sets [15]. In the process of evolution, the set of network models will be expanded, and the fitness of these network models is given by their accuracy on the verification set. In the process, two models will be randomly selected, the bad one will be eliminated directly, and the good one will become the parent node. The child nodes are formed by mutation. The child nodes are trained and verified to be put into the collection.

Real et al. [16] proposed that aging evolution, a variant of tournament selection, makes evolutionary selection tend to be a “younger” model, which can help better exploration. In addition, they compared reinforcement learning, evolutionary algorithms and random search, and found that reinforcement learning and evolutionary algorithms performed well in terms of accuracy. Compared with reinforcement learning, evolutionary algorithm can search faster and get smaller models.

## 4.3 Gradient-Based Method

The methods based on reinforcement learning and evolutionary algorithms mentioned above are essentially searching in discrete space. Scholars of CMU and Google proposed the darts method [17]. Nodes represent implicit representations, and the directed edges of connected nodes represent operator operations. The most critical trick in the darts method is to mix the candidate operations using the softmax function. Another gradient-based approach is proposed in papers published by China University of Science and Technology and Microsoft [18]. Its approach is to first embed the network structure into a continuous space in which each point corresponds to a network structure.

In order to apply the method of hyperparameter automatic tuning in NAS to other fields, the most important thing is the definition and coding of search space, which is often domain-related. In the future, as long as the data is provided, from data enhancement, to optimizer, to network structure, and then to training parameters, it can be completed automatically.

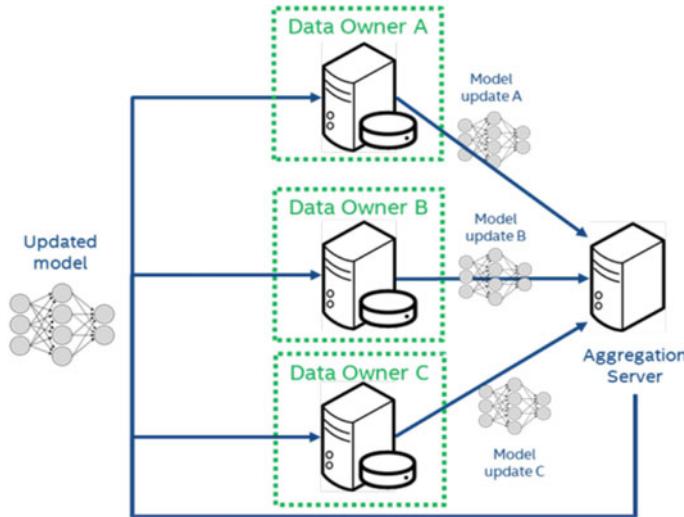
## 5 Federal Learning for Medical Image Processing

Google first proposed a federated learning system for mobile devices in 2016 [19]. The system allows users to form a consortium to train to get a centralized model, while user data is safely stored locally, which solves the problems of data privacy and security protection. As a federated learning simulation platform, the target application scene of this platform is acoustic model training [20]. This is the first attempt to apply FL technology to speech recognition tasks. Wang et al. [21] proposed federated transfer learning to solve the limitations of existing federated learning methods, and used transfer learning to provide solutions for samples and feature spaces under the framework of federated learning. Peng et al. [22] focused on introducing domain adaptation in direct transfer learning into federation learning, and proposed a federated adversarial domain adaptation method which uses adversarial techniques to solve the problem of domain transfer in federation learning. Wu et al. [23] proposed a differential privacy federated multitask learning method for effective parameter transmission with differential privacy to protect the gradient at the client level. Cao et al. [24] proposed FLTrust to provide trust for federal learning, they use the server itself to collect a clean, small training data set for the learning task, and then maintain the server model based on it to guide trust. Cao et al. [25] proposed an algorithm that learns multiple global models, each of which is learned using a randomly selected client subset. They say that the tags of the test examples predicted by their integrated global model are not affected by a limited number of malicious clients. Wang et al. [26] studied the distributed machine learning under the general Byzantine failure model, in which the Byzantine workers can arbitrarily modify the information transmitted from itself to the host.

The goal is to develop efficient distributed machine learning methods and provide provable performance assurance. WeBank AI team released Federated AI Technology Enabler [27], to provide a secure computing framework based on data privacy protection, which provides strong secure computing support for machine learning, deep learning and transfer learning algorithms. To promote medical research, protect data privacy and improve patients' brain tumor recognition results, NVIDIA, together with King's College London, launched the first federal learning system with privacy protection for medical image analysis [28]. Federal learning can achieve collaborative and decentralized neural network training without sharing patient data.

If Deep Learning is to be applied to semantic image segmentation in the field of medical images, a lot of training data is needed. Sheller et al. [29] proposed to apply Federated Learning to build an effective segmentation model on BraTS data. The experimental results show that the model of joint semantic segmentation and the model of sharing training data are similar in the final result. As shown in Fig. 3.

The goal of federated learning is to minimize the barrier with the target performance while ensuring the computing speed under the specified computing performance limits. FedSMB can achieve the accuracy of centralized training in NON-IID [30], but the number of rounds has increased. To deal with the problem of increasing the number of rounds, the federated multi-small batch, batch size is decoupled from



**Fig. 3** System architecture of federated learning

the batch count, and provides a trade-off between accuracy and communication efficiency in non-iid settings. Sui et al. [31] proposed a more efficient transfer learning scheme based on Ensemble Distillation. They first averaged the prediction probability of each teacher model for each category, and then normalized it to probability distribution using softmax. In clinical deployment, if the model trained in joint learning is applied to hospitals that are completely invisible outside the alliance, it will still suffer from performance degradation. Liu et al. [32] pointed out and solved a new problem setting of federation domain generalization, whose purpose is to learn the federation model from multiple distributed source domains, so that it can be directly extended to invisible target domains.

## 6 Conclusion

This paper summarizes various methods based on deep learning that have emerged in recent years, including image deblurring algorithms, automatic data tags, and automated network architecture design. These methods are expected to solve some difficult problems in the field of medical images, for example, the cost of high-quality training data is high, and the network structure is becoming more and more complex. Finally, the research status of federation learning and federation transfer learning is summarized. In a word, these methods can be better applied in future research.

**Acknowledgements** This research is supported by: (1) 2020-2022 National Natural Science Foundation of China under Grand (Youth) No. 52001039. (2) 2020-2022 Funding of Shandong Natural Science Foundation in China No. ZR2019LZH005.

## References

1. Seungjun, N., Tae-Hyun, K., Kyoung-Mu, L.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: International Conference on Computer Vision and Pattern Recognition. IEEE Computer Society (2016)
2. Kupyn, O., Budzan, V., Mykhailych, M.: DeblurGAN: blind motion deblurring using conditional adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2018)
3. Kai, Z., Wangmeng, Z., Shuhang, G., et al.: Learning deep CNN denoiser prior for image restoration. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017)
4. Hongyun, G., Xin, T., Xiaoyong, S., Jiaya, J.: Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2019)
5. Jiawei, Z., Jinshan, P., Jimmy, R., et al.: Dynamic scene deblurring using spatially variant recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 2521–2529. IEEE (2019)
6. Mahdi, S.H., Kostantinos, N.P.: Convolutional deblurring for natural imaging. IEEE Trans. Image Process. (2019)
7. Fangneng, Z., Hongyuan, Z., Shijian, L.: Spatial Fusion GAN for Image Synthesis. In: Internaltional Conference on Computer Vision and Pattern Recogintion (2019)
8. Tal, R.: Synthetic data and artificial neural networks for natural scene and text recognition. In: CVPR (2014)
9. Zongwei, Z., Jae, S. et al.: Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2017)
10. Chieh-Chi, K., Teng-Yok, L., Pradeep, S., Ming-Yu, L.: Localization-aware active learning for object detection (2019)
11. Donggeun, Y., In-So, K.: Learning loss for active learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2019)
12. Kaiming, H., Haoqi, F., Yuxin, W., et al.: Momentum contrast for unsupervised visual representation learning (2019)
13. Bowen, B., Otkrist, G., Nikhil, N., Ramesh, R.: Designing neural network architectures using reinforcement learning (2016)
14. Enzo, L.E.: Neural architecture search with reinforcement learning, science of the total environment (2019)
15. Esteban, R., Sherry, M., Andrew, S.: Large-scale evolution of image classifiers (2017)
16. Esteban, R., Alok, A., Yanping, H.: Regularized evolution for image classifier architecture search. In: Proceedings of the AAAI Conference on Artificial Intelligence 2018, vol. 33 (2018)
17. Hanxiao, L., Karen, S., Yiming, Y.: DARTS: differentiable architecture search (2018)
18. Renqian, L., Fei, T., Tao, Q., Enhong, C., Tieyan, L.: Neural architecture optimization (2018)
19. Jakub, K., Brenden, M., Felix, X.Y., et al.: Federated learning: strategies for improving communication efficiency (2016)
20. Dimitrios, D., Kenichi, K., Robert, G., et al.: Federated transfer learning with dynamic gradient aggregation (2020)
21. Yang, L., Yan, K., Chaoping, X., Tianjian, C., Qiang, Y.: A secure federated transfer learning framework. Intell. Syst. **35**, 78–82 (2020)
22. Xingchao, P., Zijun, H., Yizhe, Z., Kate, S.: Federated adversarial domain adaptation (2019)
23. Huiwen, W., Cen, C., Li, W.: A theoretical perspective on differentially private federated multi-task learning (2020)
24. Xiaoyu, C., Minghong, F., Jia, L., Neil, G.: FLTrust: byzantine-robust federated learning via trust bootstrapping (2020)
25. Xiaoyu, C., Jinyuan, J., Neil, G.: Provably secure federated learning against malicious clients (2021)

26. Liping, L., Wei, X., Tianyi, C., Georgios, B.G., Qing, L.: RSA: byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets (2018)
27. WeBank's AI team opened up the federated learning framework and released the federal learning white paper, <https://cloud.tencent.com/developer/article/1517805>. Last accessed 08 Oct 2019
28. NVIDIA launches the first privacy-preserving federal learning system for medical images, <http://nvidia.zhidx.com/content-6-1584.html>. Last accessed 14 Oct 2019
29. Micah, J.S., Anthony, R., Brandon, E., et al.: Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. Springer (2018)
30. Mohammad, B., Reza, N., Reihaneh, T., et al.: Federated multi-mini-batch: an efficient training approach to federated learning in non-IID environments (2020)
31. Dianbo, S., Yubo, C., Jun, Z., Yantao, J., Yuantao, X., Weijian, S.: FedED: federated learning via ensemble distillation for medical relation extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020)
32. Quande, L., Cheng, C., Jing, Q., Qi, D., Pheng-Ann, H.: FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space (2021)

# Semi-supervised Generative Adversarial Network for Face Anti-spoofing



Junting Chen , Jiwen Dong , Qingtao Hou , Shenyuan Li ,  
Xizhan Gao , and Sijie Niu

**Abstract** For the sake of safety, face recognition system often needs to include face anti-spoofing function and it has become one of the most popular topics nowadays. Traditional face anti-spoofing algorithms tend to choose hand-crafted features which are hard to cope with the changing application scenarios. Meanwhile, most state-of-the-art methods design complex neural network structure to solve the problem, which contains a large number of parameters and makes it complicated to apply to face recognition systems that need a real-time response. Moreover, the lack of training data is also a problem faced by now. To cope with the above problems, we propose a semi-supervised generative adversarial network for face anti-spoofing. Specifically, to eliminate the effect by outliers, we design a representative frame selection module to remove outliers in training video. Second, a single-frame based face anti-spoofing algorithm by using generative adversarial network (GAN) is proposed to guide the feature selection of neural network better with data augmentation by generator to solve the problem of the lack of training data. Finally, we finetune the discriminator to further improve the accuracy of classification. Our experiments are based on all four protocols of the OULU-NPU dataset. The experimental results show that our method can achieve better performance than most hand-crafted features and is competitive with recent deep learning methods.

**Keywords** Face anti-spoofing · Representative frame selection · Generative adversarial network · Deep neural network

---

J. Chen · J. Dong · S. Li · X. Gao · S. Niu

School of Information Science and Engineering, University of Jinan, Jinan 250022, China  
e-mail: [ise\\_dongjw@ujn.edu.cn](mailto:ise_dongjw@ujn.edu.cn)

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022,  
China

Q. Hou  
Ji Nan Chao Feng Intelligent Technology Co., Ltd., Jinan, China

## 1 Introduction

In order to ensure the safety of people's privacy, face anti-spoofing algorithm is playing an important role nowadays. Compared with fingerprint, iris and other features that are hard to get, face, as a biometric feature with advantages of strong anti-interference and excellent discrimination ability, has been widely used in the field of personal identity authentication. By 2014, with the popularity of convolutional neural network, this method is widely used in the face recognition field. For example, Deepface [1] proposed by Facebook AI research institute applied the deep neural network to LFW [2] dataset, has achieved 97.35% recognition accuracy, and facenet [3] proposed in 2015 has promoted the test accuracy of LFW dataset to 99.63% with the help of millions of training data.

However, while face recognition technology is developing rapidly, some practical problems existing in the face recognition technology itself have also attracted people's attention. Presentation attacks, for example, print attack, replay attack and mask attack pose a great threat to user's privacy and property security.

Face presentation attack detection (PAD), which is also called face anti-spoofing (FAS), is such a method to verify whether the user is himself or not. In the last few years, several methods by using hand-crafted features [4–7] and neural network features [8, 9, 11] have been proposed. Traditional hand-crafted based methods often tried to find invariant features to distinguish real and fake face as robustly as possible. However, because the features designed in this way are based on the subjective judgment of the designers, it is difficult to avoid the problem that the selected features are hard to describe the complex and changeable real application scenarios even through the feature fusion methods. On the other hand, features extracted by neural network mostly learn the semantic information of real and fake face for face anti-spoofing, but it often depends on the diversity and quantity of training samples, which is prone to be overfitted. Hence, how to reduce the overfitting problem of deep learning is worth exploring.

Compared with single-frame methods, multi-frame methods can often achieve better performance such as motion [10], rPPG [11] and so on, but single-frame methods could always respond more quickly than multi-frame methods in real world application scenarios. Thus, single-frame methods are more suitable than multi-frame methods to be applied in the real world.

FAS is essentially a binary classification task to classify the true face and fake face. Most existing neural network-based methods for the task of classification are based on VGG [12], ResNet [13], DenseNet [14] and so on. Generative adversarial network (GAN) [15], as a popular method, is often more concerned about the generation ability of its generator, and it has not been widely used in the field of classification. Meanwhile, because the training of GAN is a process of confrontation, with the improvement of the generator, the discriminator would also be improved. The research of CGAN [16] also proves that the generation ability of GAN can be improved effectively by adding label information. **Motivated by this, we believe that the application of GAN to the field of FAS field is worth exploring.**

Based on the discussion above, we propose a single-frame based face anti-spoofing algorithm by using generative adversarial network. Our contribution mainly includes the following points:

- (1) A method based on representative frame selection is proposed to remove outliers in training video.
- (2) A semi-supervised GAN method is proposed to explore the feasibility of applying GAN to the field of FAS.
- (3) In order to make the real data occupy more weight than the generated data in the discriminator, finetune is used to further improve the accuracy of classification.

## 2 Methodology

### 2.1 Data Pre-processing

In order to reduce the impact of background information outside the face region, we use a python package called face\_recognition [17] as our face detection method to cut the face area of each frame in the video, so as to reduce the overall size of the input data and improve the anti-interference ability. In order to make the generator extract features better, several tricks for data augmentation, such as random crop and the random horizontal flip are used for the input data.

**Representative frame selection.** Experiments show that problems such as black screen, overexposure, face detection failure occasionally appear in some frames of the original dataset, those exceptions which are treated as outliers, have a negative effect on the training process. Thus, we propose a method based on representative frame selection to remove the outliers in training video.

Through observation, it is easy to find that, for all public face anti-spoofing datasets, the normal frames in the same video usually have two properties (1) occupy a large proportion in quantity; (2) the duration of the video is usually short, so different frames of the same video usually show little diversity.

Hence there are two steps of our representative frame selection algorithm. First, find a feature map that has the overall feature of the entire video, then compare this feature map with each video frame, and select N frames with the smallest distance between frames and the feature map. The whole process can be formulated as follow:

$$a = \frac{\sum_{i=1}^n b_i}{n}, \quad (1)$$

$$b_{rep_{1..N}} = \min \left( \sum_{i=1}^n (b_i - a)^2 \right) \quad (2)$$

where  $a$  is a feature map obtained by averaging the pixel values of all video frames,  $b_{1,\dots,n}$  represents  $n$  frames and those  $n$  frames constitute a video,  $b_{rep_{1\dots N}}$  is  $N$  representative frame of this video where  $N$  could be selected according to the demand. In our experiments, we set the value of  $N$  to 30.

## 2.2 Semi-supervised Generative Adversarial Network

**Semi-supervised generative adversarial network model.** Unlike most traditional neural network methods, the classical generative adversarial network [15] is an unsupervised neural network method, which consists of two parts: generator and discriminator. Generator is used to generate random samples as close to the real samples as possible through the input random noise, while discriminator is a classifier for fake samples generated by the generator and real samples from real data, it is used to distinguish whether the input sample is from the real original data or is from the fake data which is generated by the generator. Generally speaking, the goal of generator is to generate random samples which are close to the real samples as much as possible to deceive the discriminator, while the goal of discriminator is to decide whether the input data is fake samples from generator or real samples from original data, which constitutes a process of confrontation.

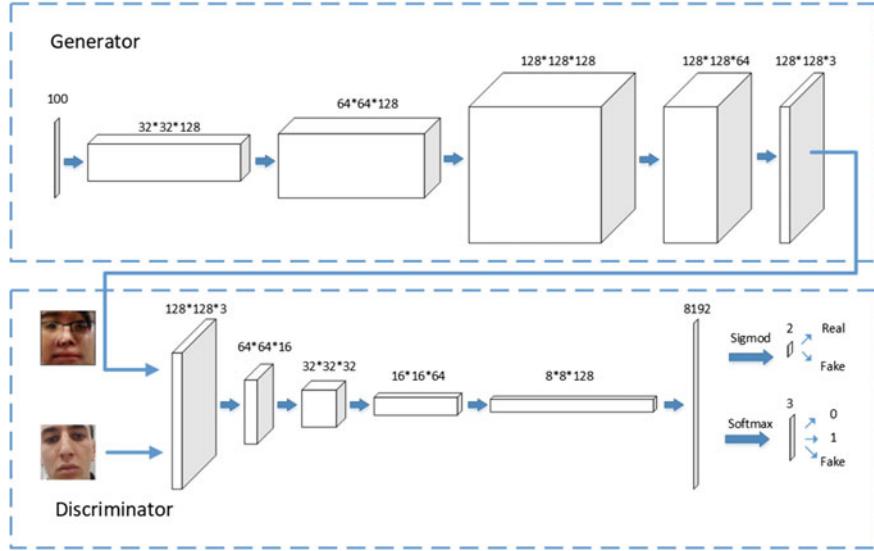
The training process of the generative adversarial network is an alternating process. First, the discriminator is trained, then the model parameters of the discriminator are fixed while the generator is training, and then the parameters of the generator are fixed during the training process of the discriminator, and so on. The loss function of the classical generative adversarial network is shown as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (3)$$

where  $G$  represents the generator,  $D$  represents the discriminator,  $x$  is original data,  $z$  is random noise. Implement details and proving process can be referred to the original paper [15].

As can be seen from the above, the original generative adversarial network method is an unsupervised process, so it can't be directly used for classification tasks. In 2016, SGAN [18] proposed an idea of adjusting the discriminator structure to force the discriminator to classify, the purpose of this idea is to improve the generation ability of the generator. Based on this idea, we modified the discriminator based on DCGAN [19] for our FAS algorithm, On the basis of the original two output neurons nodes, three output neurons nodes for classification are added to enhance the discrimination ability of the discriminator. The modified network structure is shown in Fig. 1.

Same as the original generative adversarial network, generator and discriminator are trained alternately. For the generator, the input is a random noise with the size of 100, and the output is a  $128 \times 128$  RGB image. The loss of generator is the same as that



**Fig. 1** The flow of the semi-supervised GAN

of the classical generative adversarial network. For discriminator, the input is a  $128 \times 128$  RGB image, and the output is composed of two components: discrimination loss and classification loss. Purpose of discrimination loss is to evaluate the generation quality of the generator, which is the same as the original generator loss, while the purpose of classification loss is to force the discriminator to classify the real samples by adjusting the number of output nodes. In our task, the number of output nodes, which can also be called categories is 3 (real, fake, generate image). We use Cross-entropy to do classification. The loss function is as follows:

$$L = \frac{1}{N} \sum_i - \sum_{c=1}^M y_{ic} \log p_{ic} \quad (4)$$

where  $M$  is the number of categories,  $y_{ic}$  is an indicator to determine whether the predicted label is the same as ground truth,  $p_{ic}$  is the probability where the label of sample  $i$  is  $c$ . The details of network structure are shown in Table 1.

**Finetune the discriminator.** During experiments, we found that it is even harder to train the generative adversarial network due to the addition of label information, and the result is unsatisfactory, but at this time, the discriminator has learned a lot from both unlabeled fake samples and labeled real samples. In order to make the discriminator pay more attention to the real image rather than the generated image, we finetune the current discriminator model by labeled real samples only.

**Table 1** Network architecture of DCGAN based SGAN

No	Operation	Kernel	Feature Map	BN	Stride	Activation Function	
G	1	input	—	100	—	—	
	2	FC	—	—	—	—	
	3	Reshape	—	32 * 32 * 128	Yes	—	
	4	UpSampling	2 * 2	64 * 64 * 128	—	—	
	5	Conv2d	3 * 3	64*64*128	Yes	1	LeakyReLU
	6	UpSampling	2 * 2	128 * 128 * 128	—	—	—
	7	Conv2d	3*3	128 * 128 * 64	Yes	1	LeakyReLU
	8	Conv2d	3 * 3	128 * 128 * 3	—	1	Tanh
	9	G-output	—	128 * 128 * 3	—	—	—
	10	D-input	—	128 * 128 * 3	—	—	—
	11	Conv2d	3 * 3	64 * 64 * 16	—	2	LeakyReLU
	12	Conv2d	3 * 3	32*32*32	Yes	2	LeakyReLU
D	13	Conv2d	3 * 3	16 * 16 * 64	Yes	2	LeakyReLU
	14	Conv2d	3 * 3	8 * 8 * 128	Yes	2	LeakyReLU
	15	FC1	—	2	—	—	Sigmoid
	16	FC2	—	3	—	—	softmax

### 3 Experimental Results

#### 3.1 Dataset and Metrics

The well-known OULU-NPU dataset [20] is used to evaluate our method. OULU-NPU is a dataset consisting of print attack and replay attack with high resolution videos, it contains four different protocols, and a total of 14 groups of experiments to measure the effect of illumination variation, presentation attack in structures (PAI) variation, and camera device variation. Protocol 1 focus on illumination, protocol 2 focus on PAI, and protocol 3 focus on external device. The most challenging protocol is Protocol 4, it includes all the variations in protocol 1, 2, and 3.

Furthermore, we choose APCER (Attack Presentation Classification Error Rate), BPCER (Bona fide Presentation Classification Error Rate) and ACER (Average Classification Error Rate) as our evaluation criteria of the results. Details of the definition of APCER, BPCER and ACER can be found in [21].

#### 3.2 Implementation Details

Our experiments are implemented by Pytorch. For the training phase, we chose Adam as the optimization algorithm, and the initial learning rates of generator and

discriminator are 1e-4 and 1e-3 respectively. We trained 500 epochs on RTXTitan GPU. The size of input image is  $128 \times 128$  and the batch size is 128. In the fine-tuning phase, the initial learning rate of the discriminator is 1e-3, and the learning rate is attenuated every 200 epochs with the gamma of 0.5, the number of epochs is 1000, and the batch size is 256.

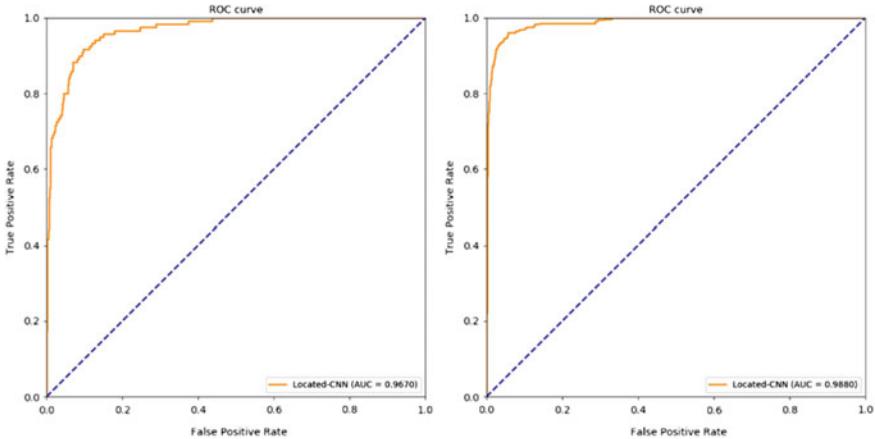
### 3.3 Comparison with Existing Methods

As shown in Table 2, our method has achieved better results than traditional hand-crafted features [5–7], and as for deep learning method Recod [22] and CPqD [23],

**Table 2** The results of all protocols of OULU-NPU

Prot	Method	APCER (%)	BPCER (%)	ACER (%)
1	MBLPQ [7]	44.2	<b>3.3</b>	23.8
	Color Texture + LBP [5]	5.0	20.8	12.9
	PML [6]	11.3	9.2	10.2
	Recod [22]	3.3	13.3	8.3
	CPqD [23]	<b>2.9</b>	10.8	<b>6.9</b>
	Ours	9.8	8.3	9.1
	Color Texture + LBP [5]	22.5	6.7	14.6
	MBLPQ [7]	19.7	6.1	12.9
2	Recod [22]	15.8	4.2	10.0
	CPqD [23]	14.7	<b>3.6</b>	9.2
	PML [6]	11.4	3.9	7.6
	Ours	<b>7.5</b>	3.9	<b>5.7</b>
	MBLPQ [7]	$12.9 \pm 4.1$	$21.9 \pm 22.4$	$17.4 \pm 10.3$
	PML [6]	$15.7 \pm 21.8$	$15.8 \pm 15.4$	$15.8 \pm 15.1$
3	Color Texture + LBP [5]	$14.2 \pm 9.2$	$8.6 \pm 5.9$	$11.4 \pm 4.6$
	Recod [22]	$10.1 \pm 13.9$	$8.9 \pm 9.3$	$9.5 \pm 6.7$
	CPqD [23]	$6.8 \pm 5.6$	$8.1 \pm 6.4$	$7.4 \pm 3.3$
	Ours	<b><math>4.4 \pm 4.0</math></b>	<b><math>7.5 \pm 7.5</math></b>	<b><math>6.6 \pm 2.6</math></b>
	PML [6]	$61.7 \pm 26.4$	$13.3 \pm 13.7$	$37.5 \pm 14.1$
	MBLPQ [7]	$49.2 \pm 27.8$	$24.2 \pm 27.8$	$36.7 \pm 4.7$
4	Color Texture + LBP [5]	$29.2 \pm 37.5$	$23.3 \pm 13.3$	$26.3 \pm 16.9$
	Recod [22]	$35.0 \pm 37.5$	<b><math>10.0 \pm 4.5</math></b>	$22.5 \pm 18.2$
	CPqD [23]	$32.5 \pm 37.5$	$11.7 \pm 12.1$	$22.1 \pm 20.8$
	Ours	<b><math>9.3 \pm 4.2</math></b>	$19.8 \pm 19.0$	<b><math>14.6 \pm 11.6</math></b>

The bold values indicate the best results in this group of experiments



**Fig. 2** ROC curves of protocol 1 and 2 in OULU-NPU while the result of protocol 1 is one the left and the result of protocol 2 is on the right

our method has also obtained competitive results. Because the results of protocol 1, 2 and 3 measure the effect of illumination variation, presentation attack in structures (PAI) variation, and camera device variation respectively, the experimental results show that for PAI variation and camera device variation, our method is stable. The results of protocol 4 also show that our method has a better effect on APCER and ACER than any other methods above. Low APCER means that our method has a strong ability to reject presentation attacks, which makes our method more suitable for real-world applications.

Moreover, FAS is a task of binary classification. In Fig. 2, we evaluate the influence of different thresholds on our method by ROC curves based on protocol 1 and 2 as an example. The area under the ROC curve can be regarded as an intuitive evaluation of its results. As we can see, both curves have a good trend and ROC curve on protocol 2 achieve a better performance than the curve on protocol 1.

## 4 Conclusion and Future Work

We explore the feasibility of forcing the discriminator of the generative adversarial network to predict class labels for classification task in this paper. A method based on representative frame selection and finetune is also proposed to further improve the accuracy of classification. The experimental results show that our proposed method can achieve better results than most hand-crafted features and is competitive with some recent deep learning methods. Meanwhile, the effective application of generative adversarial network in the field of classification still needs further exploration.

**Acknowledgements** This work is supported by the National Natural Science Foundation of China under Grant No. 61872419, No. 61873324, the Natural Science Foundation of Shandong Province, China, under Grant No. ZR2020QF107, No. ZR2020MF137, No. ZR2019MF040, ZR2019MH106, No. ZR2018BF023, the China Postdoctoral Science Foundation under Grants No. 2017M612178. University Innovation Team Project of Jinan (2019GXRC015), Key Science & Technology Innovation Project of Shandong Province (2019JZZY010324,2019JZZY010448), and the Higher Educational Science and Technology Program of Jinan City under Grant with No. 2020GXRC057.

## References

1. Taigman, Y., Yang, M., Ranzato, M. A., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701–1708 (2014)
2. Gary, B. H., Marwan, M., Tamara, B., Eric, L.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition (2008)
3. Schroff, F., Dmitry, K., James, P.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
4. Ivana, C., André, A., Sébastien, M.: On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG) (2012)
5. Zinelabidine, B., Jukka, K., Abdenour, H.: Face anti-spoofing based on color texture analysis. In: 2015 IEEE International Conference on Image Processing (ICIP) (2015)
6. Bekhouche, S.E., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., Hadid, A.: Pyramid multi-level features for facial demographic estimation. *Expert Syst. Appl.* **80**, 297–310 (2017)
7. Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S. E., Hadid, A.: A competition on generalized software-based face presentation attack detection in mobile scenarios. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 688–696 (2017)
8. Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Liu, W.: Face anti-spoofing: model matters, so does data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3507–3516 (2019).
9. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5295–5305 (2020)
10. Bharadwaj, S., Dhamecha, T. I., Vatsa, M., Singh, R.: Computationally efficient face spoofing detection with motion magnification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 105–110 (2013)
11. Yaojie, L., Amin, J., Xiaoming, L.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018).
12. Karen, S., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. In arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
13. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Huang, G., Liu, Z., Van der Maaten, L., Weinberger, K. Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y.: Generative adversarial networks. arXiv preprint [arXiv:1406.2661](https://arxiv.org/abs/1406.2661). (2014)
16. Mehdi, M., Simon, O.: Conditional generative adversarial nets. arXiv preprint [arXiv:1411.1784](https://arxiv.org/abs/1411.1784) (2014)
17. Adam Geitgey., [https://github.com/ageitgey/face\\_recognition](https://github.com/ageitgey/face_recognition), Copyright (c). Last accessed 2017
18. Augustus, O.: Semi-supervised learning with generative adversarial networks. arXiv preprint [arXiv:1606.01583](https://arxiv.org/abs/1606.01583) (2016)
19. Alec, R., Luke, M., Soumith, C.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434) (2015)
20. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (2017)
21. Ivana, C., Amir, M., Andrew, A., Sébastien, M.: Evaluation methodologies for biometric presentation attack detection. In: Handbook of Biometric Anti-Spoofing, pp. 457–480. Springer, Cham (2019)
22. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint [arXiv:1602.07360](https://arxiv.org/abs/1602.07360) (2016)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

# Improving Apple Detection Using RetinaNet



Zhen Ma  and Nianqiang Li 

**Abstract** With the rapid modernization of agriculture and the increasing demand for fruits, automated fruit picking tasks are particularly important in fruit production. Computer vision-based fruit target detection is one of the key technologies. Traditional fruit detection methods are limited by the fact that the digital images captured by cameras are susceptible to light, and there may be overlap and occlusion between the fruit and the leaves, which are common situations that can greatly affect traditional fruit detection methods. With the development of deep learning techniques, many target detection techniques have emerged. To improve the accuracy and detection speed of fruit detection, this paper adopts and improves RetinaNet, using MobileNetV3 as one of the feature extraction network, which greatly reduces the inference time of detection models in embedded devices. In order to improve the detection accuracy of the small target of fruits, this paper makes some improvements to the feature extraction network and feature pyramid network in the network, and optimizes the size of anchors with a clustering algorithm. Through experiments, it is shown that the improved RetinaNet algorithm proposed in this paper has high accuracy in apple detection task and better robustness in dark light, overlapping and occlusion situations.

**Keywords** RetinaNet · MobileNetV3 · Apple detection

## 1 Introduction

China is a large fruit producing country, and the use of automated fruit picking equipment can greatly reduce human and material resources, and an important prerequisite for this mechanical task is to have accurate machine vision-based fruit detection and positioning technology. Conventional inspection methods identify fruits based on fruits' features such as shapes, colors and textures [1–4]. The detection speed of these methods is usually fast, but the pictures taken by the camera are not ideally

---

Z. Ma · N. Li (✉)  
University of Jinan, Jinan 250022, China  
e-mail: [ise\\_linq@ujn.edu.cn](mailto:ise_linq@ujn.edu.cn)

images that contain only complete fruits, and often encounter complex situations such as overlapping fruits, leaves obscuring fruits, and large differences between individual fruits, which can make traditional detection methods very difficult.

Deep learning is one of the frontier technologies in machine learning and artificial intelligence research, and deep learning techniques have brought revolutionary advances in machine learning and computer vision. Target detection refers to separating the background region from the region of interest and determining the classes and location of the region of interest in the input of an unknown image. In recent years, due to the breakthrough of deep learning technology in target detection [5, 6], there have been many scholars applying this technology to fruit detection. The current target detection methods based on deep convolutional networks are mainly divided into two categories, one category is two-stage detection algorithms such as RCNN [7], Fast-RCNN [8], Faster-RCNN [9], Mask-RCNN [10]; the other category is one-stage detection algorithms such as SSD [11], YOLO [12–14], RetinaNet [15].

Bargoti et al. [16] used Fast-RCNN network for fruit detection, and to reduce the computational effort, they split the original high-resolution image and detected each piece separately. They also used flip shift and color space transformation to enhance the dataset and reduce the occurrence of overfitting cases. However, Fast-RCNN is a two-stage detection algorithm, which has a long detection time and is not optimized for small targets. Tian et al. [17] used a modified YOLO-V3 network for apple detection, and modified YOLO-V3 using DenseNet to improve the feature extraction capability of the network.

Compared with the two-stage detection algorithms, the one-stage detection algorithms skip the region proposal stage and complete the prediction of target class and target localization simultaneously in the convolutional network, which greatly speed up the detection speed and are more suitable for real-time detection of fruits. However, one-stage detection algorithms usually use the mechanism of dense sampling of candidate regions, which can lead to the occurrence of category imbalance [18], i.e., the quantity of negative samples is much larger than the quantity of positive samples, making the training process affected and thus reducing the accuracy of target detection.

In RetinaNet, He et al. [15] proposes a new loss function, Focal Loss, which is characterized by its small impact on the loss function for easy samples and still maintains a high loss for hard samples, thus making the training process more stable and improving the efficiency and accuracy of the detector.

In summary, this paper adopts RetinaNet as the target detection framework and uses MobileNetV3 as its feature extraction network for the application scenario of this paper. In order to improve the accuracy of small target detection, this paper improves the RetinaNet network structure by making a series of modifications to the feature extraction network and FPN so that the low-level features of the image samples have better semantic information of the high-level features. The anchors parameters in the original RetinaNet are not applicable to the application scenario of this paper, so a better anchors parameter is calculated by the K-means distance algorithm to improve the detection accuracy and increase the recall.

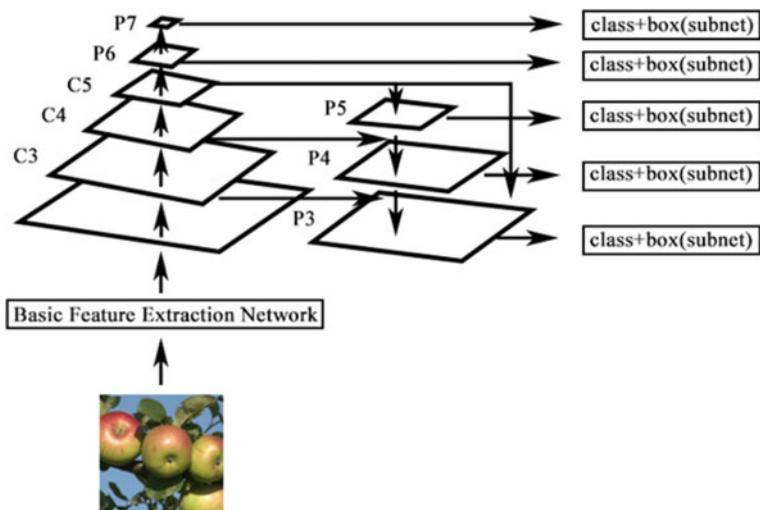
The algorithm proposed in this paper takes into account the better accuracy while completing the lightweight of the model, so that the embedded device can complete the real-time orchard apple detection in real time and be fully prepared for the subsequent fully automated fruit picking task.

## 2 RetinaNet Structure and Improvement

RetinaNet is a unified target detection network consisting of a feature extraction network, a feature pyramid network and two sub-networks. It improves the accuracy of target detection, especially in the detection of small objects. This paper improves on it, and its network structure is shown in Fig. 1. The backbone network mainly obtains the feature map of the whole input image through a series of convolutional operations. Two sub-networks classify and localize the target image to be detected based on the output of the backbone network, respectively.

### 2.1 Feature Extraction Network

MobileNets [19] is based on a streamlined architecture that uses deeply separable convolutions to build lightweight deep neural networks. The network introduces two simple global hyper parameters that effectively balance between latency and accuracy. In this paper, MobileNetV3 is used as the base feature extraction network



**Fig. 1** Overall structure of improved RetinaNet

of RetinaNet for practical application scenarios to shorten the inference time of the model and to achieve the real-time detection task of apples in orchards using embedded devices.

## 2.2 Feature Pyramid Network

Feature Pyramid Network (FPN) [20] has been a fundamental component in multi-scale target detection and can cope well with target detection tasks of different sizes. The high-level features of image samples contain rich semantic information, but it is difficult to predict the location of the target accurately due to low resolution. In contrast, the low-level features of image samples have less semantic information, so that they can accurately contain the location information of objects due to their high resolution. According to this feature, FPN fuses the feature maps of different layers, enabling better recognition of small objects. However, in the FPN module of RetinaNet, after multiple convolution and upsampling operations, the semantic information of the higher-level features of the image samples is difficult to reach the lower-level feature layers, making the lower-level features used to detect small objects lack some semantic information of the higher-level features.

In the application scenario of this paper, for the problem of low detection accuracy in the detection task of small targets such as apples, the C5 feature layer of the feature extraction network is stacked with the P3 feature layer of the FPN after 4 times up-sampling, and the stacked feature channels are compressed and fused to the original number of feature channels by a  $1 \times 1$  convolutional layer to improve the semantic information of the high-level features contained in the low-level features of the image samples.

## 2.3 Focal Loss

In the process of object detection algorithm training, there is a class imbalance problem, the most serious of which is the positive and negative sample imbalance, i.e., the number of negative samples is often larger than the number of positive samples. In some two-stage object detection methods, like Faster R-CNN, a significant portion of negative samples is first filtered out using RPN, and then a deep neural network is used to make accurate category detection and position regression for each candidate frame. In contrast, one-stage target detection methods do not distinguish between positive and negative samples in advance, and directly performs category detection and position regression on the pre-set anchors. So, in general one-stage detection algorithm is faster without high accuracy. To solve this problem, He et al. mentioned Focal Loss in the RetinaNet algorithm.

The traditional cross-entropy loss function is shown as follows:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases} \quad (1)$$

where  $y \in \{\pm 1\}$  refers to the manually labeled classes and  $p \in [0, 1]$  is the probability that the model predicts the classes  $y = 1$ .

For simplicity, we let

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \quad (2)$$

Then we can obtain the formula as follows:

$$CE(p, y) = \log(p_t) \quad (3)$$

To solve the problem of classes imbalance caused by the number of negative samples being much larger than the number of positive samples, we can introduce a weighting factor  $\alpha$ ,  $\alpha$  is defined as follows:

$$\alpha = \begin{cases} \alpha & \text{for class 1} \\ 1 - \alpha & \text{for class -1} \end{cases} \quad (4)$$

We can treat  $\alpha$  as a hyper parameter and calculate the optimal value by cross-validation method. The loss function then becomes as follows.

$$CE(p_t) = -\alpha_t \log(p_t) \quad (5)$$

In order to better distinguish between easy examples and hard examples, the detector focuses more on hard examples.

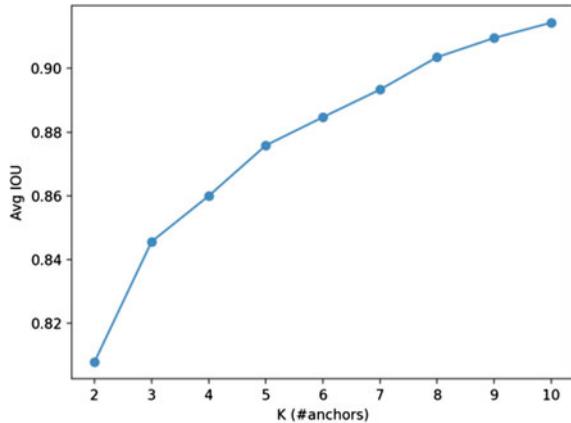
Focal Loss introduces a tunable focusing parameter ( $\gamma > 0$ ), and the final formula for Focal Loss is shown below.

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (6)$$

For the dataset used in this paper, we set  $\gamma$  to be 2 to get better results for the trained model.

## 2.4 Improving Anchor Using Clustering Algorithm

By using the anchors mechanism [9], the computational effort in the training phase is greatly reduced. Since the setting parameters of anchors can vary between different datasets, the parameters of anchors are recalculated by using the K-means clustering algorithm for the application scenario of this paper, which make the model of this

**Fig. 2** Clustering result

paper work better on the fruit dataset. In this paper, the average IOU in each case is calculated by using the K-means distance algorithm between K belonging to (2,10) anchors, and the calculation results are shown in Fig. 2.

It can be seen that the slope of the anchors-average IOU curve changes significantly when the number of anchors is 3. Therefore, when the size of 3 anchors is selected, the complexity of training can be reduced while the accuracy of the model can be taken into account. At this time. The three anchors' sizes are [27 × 27, 36 × 36, 46 × 46]. The anchors sizes obtained by clustering the ground truth bounding box are closer to the true values, which make it easier to fit the model to the true position of the target, thus reducing the training difficulty of the model.

### 3 Experimental Results and Analysis

#### 3.1 Experimental Data

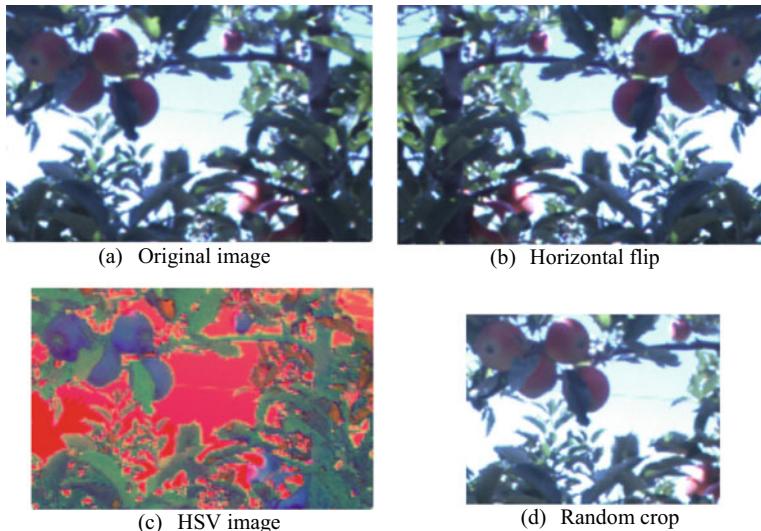
The dataset used in this paper is published by Suchet Bargoti and James Underwood in [21] and can be downloaded at [22]. This paper uses its apple dataset. The dataset provides circular annotations for the fruits, which is converted into a rectangular box representation containing four vertices to better fit the network parameters of this paper (Table 1).

#### 3.2 Data Augmentation

Since this dataset has only more than 1000 apple images, it is easy to overfit during the training process. For this reason, we used the following methods to augment the

**Table 1** Apple dataset parameters

Set	Raw image size	Image size	Number of image
train	1616 × 1232	202 × 308	896
val	1616 × 1232	202 × 308	112
test	1616 × 1232	202 × 308	112
train + val	1616 × 1232	202 × 308	1008



**Fig. 3** Data augmentation

dataset to enhance the robustness of the model. (1) Convert all images to HSV color space to enhance the contrast between foreground and background in apple images. (2) Flip all images horizontally to expand the original dataset by a factor of 2. (3) Randomly crop the images during training by randomly cropping 60–90% part and scaled to the size needed by the network. After the above operations, the dataset is expanded to 3 times of the previous size, which greatly reduces the occurrence of overfitting and enhances the generalization of the model. Figure 3 shows the images after data augmentation.

### 3.3 Evaluation Criteria

In this paper, F1-score is used as the evaluation index of the target detection model. F1-score is the harmonic average of *Recall* and Precision. The formulas of *Recall* and *Precision* are as follows:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Among them,  $TP$  is True Positives, which means that the sample is divided into positive samples and the allocation is correct.  $FP$  is  $FN$ , that is, False Negatives, which means that the sample is divided into negative samples but the allocation is wrong.  $FN$  is False Negatives, which means that the sample is divided into negative samples but the allocation is wrong. Thereby, the calculation formula of F1-score is obtained, as shown below:

$$F1 = 2 \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

The higher the F1-score, the more robust the model.

### 3.4 Experimental Results

In this paper, we conduct comparison experiments by using different detection algorithms on the same datasets, and the experimental results are shown in Table 2. By improving the RetinaNet network architecture and using MobileNetV3 as its feature extraction network, the detection speed is greatly improved, from 8 to 37FPS, and the F1-score is also improved by 20%.

Figure 4 shows the detection results of the improved algorithm for some test samples. In these test result plots, some common cases that are unfavorable to Apple detection are shown. For example, the picture in the upper left corner in Fig. 4, the cyan colored apples are extremely similar in color to the large green leaves in the background. In the top right image, the light is blocked and in dim light, a similar situation to the top left picture occurs, where the apples largely blend in with the background. In the two pictures at the bottom of Fig. 4, there is an overlap between apples and the leaves obscure the apples. Although these common unfavorable situations above can bring great impact on the target detection, however, the improved RetinaNet in this paper still plays a good effect, not only detecting the apples in the images, but also accurately labeling their positions in the images.

**Table 2** Experimental results

Model	Backbone	F1
Faster-RCNN	ResNet50	0.878
RetinaNet	ResNet50	0.865
Proposed	MobileNetV3	<b>0.946</b>

**Fig. 4** Detection result

### 3.5 Experimental Analysis

The original RetinaNet uses Focal Loss as the loss function, and its F1-score is not much different from Faster-RCNN under the premise of guaranteeing the detection speed, but as a two-stage detection algorithm, the Faster-RCNN model is relatively large and not well able to accomplish the real-time target detection task. Therefore, this paper adopts the lightweight RetinaNet detection model and uses MobileNetV3 as its feature extraction network, which is faster and can complete the real-time apple detection task. For the detection of the small target of apple, the detection accuracy is greatly improved by improving the RetinaNet network structure, which makes the semantic information of the high-level features well integrated with the low-level features, and by using the K-means clustering algorithm to calculate the size of the anchors suitable for this dataset. For complex and common unfavorable cases, such as dark light, overlap and occlusion, the algorithm in this paper still has excellent performance. However, this algorithm also has some shortcomings, such as some detection targets are missed. In the next work, we will continue to improve this model and add other datasets to further improve the robustness and accuracy of the detection model.

## References

1. Wang, Z., Walsh, K. B., Verma, B.: On-tree mango fruit size estimation using RGB-D images. *Sensors* **17**(12), 2738 (2017)
2. Payne, A.B., Walsh, K.B., Subedi, P.P., Jarvis, D.: Estimation of mango crop yield using image analysis-segmentation method. *Computers Electron Agric* **91**, 57–64 (2013)
3. Nanaa, K., Rizon, M., Abd Rahman, M.N., Ibrahim, Y., Abd Aziz, A. Z.: Detecting mango fruits by using randomized hough transform and back propagation neural network. In: 2014 18th International Conference on Information Visualisation, pp. 388–391. IEEE (2014)
4. Rizon, M., Yusri, N. A. N., Kadir, M. F. A., bin Mamat, A. R., Abd Aziz, A. Z., Nanaa, K.: Determination of mango fruit from binary image using randomized Hough transform. In: Eighth International Conference on Machine Vision (ICMV 2015), vol. 9875, pp. 987503. International Society for Optics and Photonics (2015)
5. Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.: Object detection with deep learning: a review. *IEEE Trans. Neural Networks Learn. Syst.* **30**(11), 3212–3232 (2019)
6. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: a survey. *Int. J. Comput. Vision* **128**(2), 261–318 (2020)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
8. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
13. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
14. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. (2018)
15. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the Ieee International Conference on Computer Vision, pp. 2980–2988 (2017)
16. Bargoti, S., Underwood, J.P.: Image segmentation for fruit detection and yield estimation in apple orchards. *J. Field Robot.* **34**(6), 1039–1060 (2017)
17. Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z.: Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* **157**, 417–426 (2019)
18. Oksuz, K., Cam, B. C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: a review. *IEEE Trans. Pattern Anal. Machine Intell.* (2020)
19. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Adam, H.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
20. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

21. Bargoti, S., Underwood, J.: Deep fruit detection in orchards. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 3626–3633. IEEE (2017)
22. ACFR FRUIT DATASET: <http://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/>. Last accessed 8 May 2021

# Megavoltage Computed Tomography (MVCT) Imaging Quality Improvement via Convolutional Neural Network



Zengjing Zhao , Jiwen Dong , Sijie Niu , Yan Zhang , and Jian Zhu

**Abstract** Clinical application of Megavoltage Computed Tomography (MVCT) is limited to three-dimensional visualization and position of the patient. The noise and low soft tissue contrast of MVCT make it difficult to offer accurate target volume delineation for adaptive radiotherapy and accurate image registration for image-guided radiotherapy, resulting in the less clinical application prospect of MVCT. In this work, we developed a deep learning-based approach to improve the MVCT image quality, so as to learn a mapping between MVCT images and paired planning kilovoltage CT (KVCT) images. The root mean squared error (RMSE), structural similarity (SSIM), and peak signal-to-noise ratio (PSNR) were used to quantify. Specifically, the PSNR and SSIM of the processed images improved to  $27.94 \pm 1.64$  dB and  $0.93 \pm 0.01$  compared with  $24.67 \pm 1.16$  dB and  $0.88 \pm 0.02$  in the original MVCT images, and the RMSE decreased from  $23.5 \pm 3.14$  HU to  $16.34 \pm 3.39$  HU. The processed images were obtained using the proposed method with less noise and higher soft tissue contrast in comparison with machine learning-based method. The processed MVCT increases the feasibility of depicting the anatomical delineation of tumor and organs at risk, and therefore enables the promising application of image-guided and adaptive helical TomoTherapy.

**Keywords** MVCT · KVCT · Convolutional neural network (CNN)

---

Z. Zhao · J. Dong · S. Niu

School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

Z. Zhao · Y. Zhang · J. Zhu ()

Department of Radiation Oncology Physics and Technology, Shandong Cancer Hospital affiliated to Shandong First Medical University, Jinan 250022, China

e-mail: [zhujian@sdfmu.edu.cn](mailto:zhujian@sdfmu.edu.cn)

## 1 Introduction

Helical TomoTherapy (HT) is a revolutionary and neoteric technique to radiotherapy, including an on-board imaging system that can be used to obtain MVCT images of the patient before the treatment [1, 2]. After the MVCT acquisition, an KVCT-to-MVCT image registration was registered for more accurate localization of patients prior. In addition, MVCT images have also been used to monitor the position error of the patient during the radiotherapy and therefore the treatment can be adapted perfectly [3, 4]. The on-board imaging system uses Megavoltage (MV) imaging does not require additional hardware and a separate imaging isocenter, where the imaging beam and the treatment beam are identical with same isocenter (treatment mode and imaging mode are 6 MV and 6 MV X-ray respectively [5]). At this higher energy spectrum, compton scattering dominates the photon interactions in all body tissues and streaking artifacts and the attenuation is proportional to the tissue density, which results in the amplification of doped noise [6]. The low images quality of MVCT limits its application potential in radiotherapy.

In recent years, some scholars had tried to use image reconstruction and image post-processing methods to denoise MVCT images, aiming to improve its images quality and increase soft tissue information. For instance, by reconstructing MVCT images with reduced noise and discernible boundary, improving MVCT images quality by using tensor framework [7]. Yet this method does not significantly improve the contrast of soft tissues, and causes the resolution to be reduced. An anisotropic diffusion filter has ever been used to improve deformable registration of the MVCT images [8]. However, this method typically small feature contrast was reduced [9]. An algorithm that combines block matching 3D (BM3D) filtering and saliency map was applied to MVCT images denoising and improving soft tissue information [9]. However, this algorithm needs to create a saliency map based on general experience. A method of combining BM3D filtering and discriminative feature representation (BM3D + DFR) was applied to MVCT images denoising [10], which has improved the quality of MVCT image. While it takes too much time and is not convenient for clinical use. Recently, the generative adversarial network (GAN) has been broadly used for image-to-image translation [11–13], Especially in the medical field [14, 15]. A CycleGAN [16] architecture improves the image quality of MVCT through end-to-end learning from MVCT to KVCT [17]. Although the method can learn the translation mappings in the absence of aligned paired images, it may generate non-existent objects or features.

In this study, we apply a convolutional neural networks (CNN) on MVCT images denoising. The CNN model is based on convolutional encoding-decoding [18] framework, which includes deconvolution network and shortcut connections [19]. The method not only makes running speed meet the clinical requirements, but also correct the CT numbers in MVCT images by accurate Hounsfield Units (HU) information from previous planning KVCT, so as to reduce the noise. Besides that, we use overlapped patches in CT images, in this way, not only achieve the effect of data enhancement, but also detect local differences in perception [20].

## 2 Materials and Methods

### 2.1 Data Acquisition and Image Processing

Head and neck (H&N) patients from Shandong Cancer Hospital and Institute, of which the MVCT and KVCT images were collected, were divided into 34 training (1484 transversal MVCT slices) and 4 testing (263 transversal MVCT slices) patients. MVCT images were acquired by the helical TomoTherapy system with voxel spacing of  $0.76 \times 0.76 \times 2.00 \text{ mm}^3$  and image size of  $512 \times 512$ . The KVCT images were acquired by Brilliance Big Bore CT scanner with voxel spacing of  $1.06 \times 1.06 \times 3.00 \text{ mm}^3$  and imaging matrix size of  $512 \times 512$ . For training, we matched KVCT to MVCT to generate paired KVCT images as the ground truth though the opensource image registration toolbox (Elastix) [21]. The original KVCT images were resampled to the same resolution of  $0.76 \times 0.76 \times 2.00 \text{ mm}^3$ , the same as MVCT.

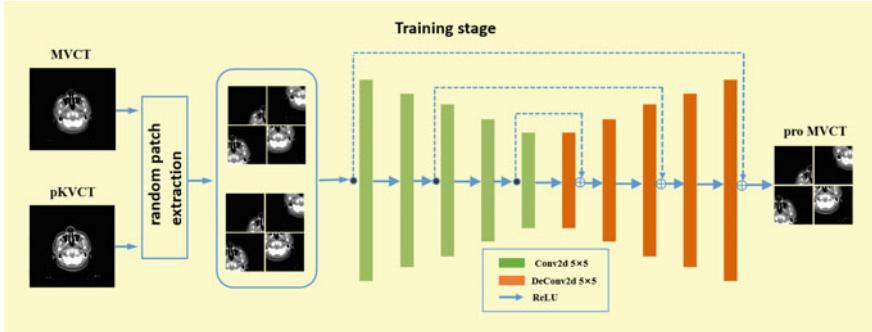
### 2.2 Network Design

The present study used a residual encoder-decoder convolutional neural network architecture. This framework has been successfully used in image restoration such as denoising and super-resolution [16, 19], and here this is the first application on the MVCT denoising.

Denoising is considered a “low-level” task, because there is no feature extraction, the level of most image denoising models is limited. Therefore, the network has only 10 layers, of which 5 layers are convolutional layers and 5 symmetrically arranged deconvolutional layers. The convolutional layer is used to remove noise in the MVCT images, the deconvolutional layer is used to reconstruct the MVCT images from the extracted features, and shortcuts connected are added to improve the learning process of the network. Each layer includes a convolution operation and rectified linear units (ReLU). The shortcuts connected occurs before the ReLU. The number of feature maps in the last layer was 1, and other layers was 96. All strides of convolution operation were set to 1, and padding was set to 0. We extracted patches of the same size from MVCT and corresponding KVCT images. The schematic flow chart is shown in Fig. 1.

The loss function used the mean squared error (MSE), which measures the pixel-wise difference between processed MVCT images  $I_{proMVCT}$  and the paired KVCT images  $I_{pKVCT}$ . The formula of MSE is shown in Eq. (1).

$$Loss_{MSE} = \frac{1}{N} \sum_{i=1}^N \|I_{proMVCT} - I_{pKVCT}^2\| \quad (1)$$



**Fig. 1** Schematic flow chart of the proposed method. pKVCT, paired KVCT; pro MVCT, processed MVCT

### 2.3 Metrics for Evaluation

In this work, the PSNR, SSIM, and RMSE were used to quantify the correction accuracy. We used Two-tailed paired t-tests (95% confidence interval) to verify the statistical significance, and recorded the P-values and t-statistics.

RMSE is a frequently used measurement criteria between the values predicted by a model and the values observed. It is defined as

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M |f(i, j) - t(i, j)|^2} \quad (2)$$

where  $f(i, j)$  is the value of pixel  $(i, j)$  in the KVCT image,  $t(i, j)$  is the value of pixel  $(i, j)$  in the MVCT image, and  $M$  is the total number of pixels.

In the field of image denoising, PSNR is an authoritative and commonly used index to evaluate image quality. It is defined as

$$PSNR = 10 \times \log_{10} \left( \frac{MAX^2}{MSE} \right) \quad (3)$$

where  $MAX$  is the possible maximum signal intensity, and  $MSE$  is the mean-squared error of the image.

SSIM is an indicator of the degree of similarity between the two images quantizing. It is defined as

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\delta_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\delta_x^2 + \delta_y^2 + c_2)} \quad (4)$$

where  $\mu_x$  and  $\mu_y$  are the mean value of KVCT and sKVCT image.  $\delta_x$  and  $\delta_y$  are the standard deviations of KVCT and sKVCT image, and  $\mu_{xy}$  is the covariance of two

images of KVCT and sKVCT. The parameters  $c_1 = (k_1 L)^2$  and  $c_2 = (k_2 L)^2$  are usually fixed values, where  $k_1 = 0.01$  and  $k_2 = 0.02$ .  $L$  is the range of pixel values.

### 3 Experimental Results

#### 3.1 Network Training

MVCT and KVCT images are normalized to the range of  $(-1, 1)$ , which can accelerate and stabilize training convergence. The patch size was set to  $128 \times 128$  as input. And beyond that, the rotation and flipping (vertical and horizontal) were used for data augmentation. The network was implemented using Pytorch, and trained using an NVIDIA GeForce GTX 1080 Ti GPU. In the training stage, adam optimizer was used to train the model. In 200 epochs, the learning rate was set to 10 $-4$  at the beginning, and slowly dropped to 10 $-5$  in the process. The training images were done on patches of  $128 \times 128$  sizes. When testing, the images size of  $512 \times 512$  were used as input.

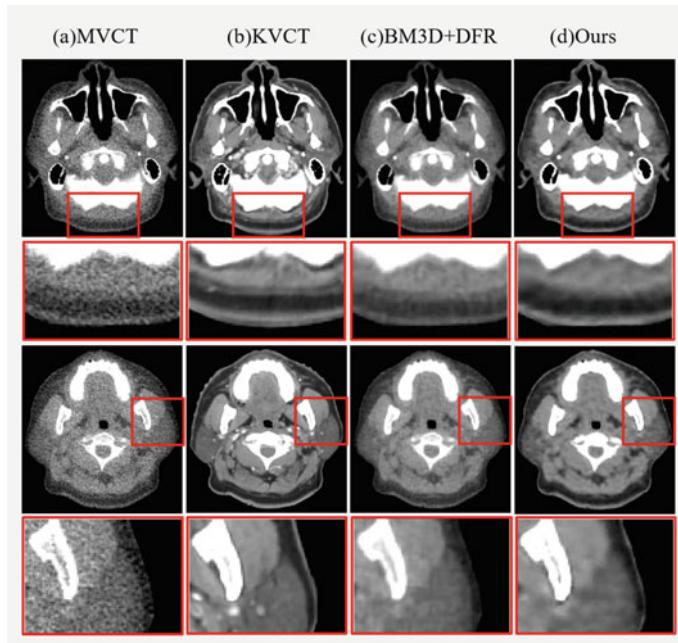
#### 3.2 Qualitative Results

The original MVCT, KVCT, and the post-processed MVCT images were shown and compared with that of using BM3D + DFR method. Figure 2c, d shows a significant noise reduction compared to the original MVCT (Fig. 2a). Compared with the BM3D + DFR method, the processed images obtained by our method has less noise, where the red box represents the enlarged soft tissue.

#### 3.3 Quantitative Analysis

The statistical evaluation results of various indicators are shown in Table 1. Compared with the BM3D + DFR, our method has achieved better results, the difference between both methods was statistically significant ( $p < 0.001$ ).

Moreover, transversal slices from one testing patient are shown in Fig. 3. Figure 3a shows that HU values of the processed MVCT by our method are similar to that of KVCT, where the orange line marked across the soft tissue and bone areas. Figure 3(b) shows that HU values of original MVCT are saltant, while HU values of processed MVCT are smoothed and similar that of KVCT, where the blue line marked only across the soft tissue area.



**Fig. 2** Head and neck images of original MVCT (a), KVCT (b) and the processed MVCT by BM3D + DFR (c) and our method (d). Displaying window is (40,400) HU

**Table 1** The PSNR, SSIM and RMSE calculation by BM3D + DFR and our methods from all 4 testing patients (263 transversal slices)

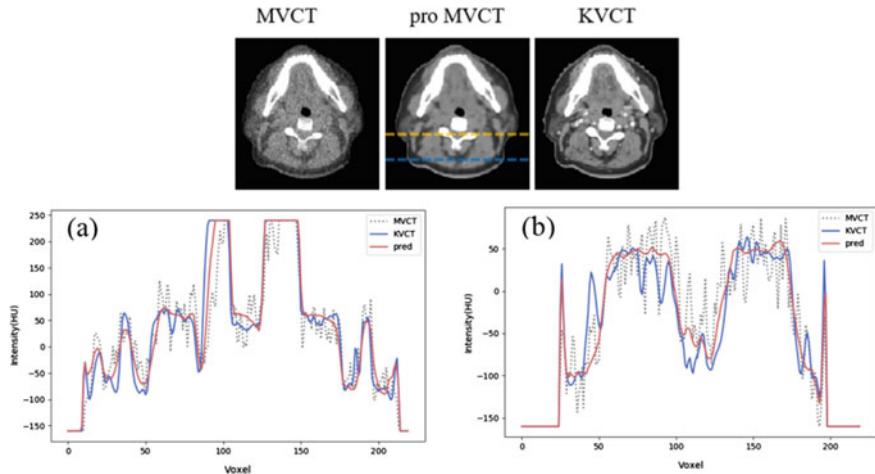
	MVCT (A)	BM3D + DFR (B)	Ours (C)	A versus C*	B versus C*
PSNR	$24.67 \pm 1.16$	$25.16 \pm 1.38$	<b><math>27.94 \pm 1.64</math></b>	$t = -41.151 p < 0.001$	$t = -35.296 p < 0.001$
SSIM	$0.88 \pm 0.02$	$0.90 \pm 0.01$	<b><math>0.93 \pm 0.01</math></b>	$t = -43.852 p < 0.001$	$t = -35.379 p < 0.001$
RMSE	$23.55 \pm 3.14$	$22.35 \pm 3.65$	<b><math>16.34 \pm 3.39</math></b>	$t = -76.941 p < 0.001$	$t = -53.827 p < 0.001$

The bold values represent better than the original MVCT images and the MVCT images processed by the BM3D+DFR

\* t and p value from two-tailed paired t-tests

## 4 Conclusion and Future Work

In this article, we used paired KVCT and MVCT for end-to-end learning to improve the quality of MVCT images. The processed images had less noise and higher soft tissue contrast in comparison with original MVCT images. Our method runs faster than machine learning-based method (BM3D + DFR) to better meet the clinical



**Fig. 3** HU line profiles of MVCT, KVCT, and processed MVCT

needs. With the integration of it in the adiotherapy, the method may bring the potential prospect in Helical TomoTherapy. In order to better apply it to the clinic, how to make the edge clearer on the premise of improving the contrast is our next research work.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (grant numbers 81671785, 81530060 and 81874224), the Shandong Provincial Natural Science Foundation (ZR2016HQ09, ZR2020LZL001), and the Academic promotion program of Shandong First Medical University (2020RC003, 2019LJ004).

## References

1. Gupta, T., Upasani, M., Master, Z., Patil, A., Phurailatpam, R., Nojin, S., Kannan, S., Godasastri, J., Jalali, R.: Assessment of three-dimensional set-up errors using megavoltage computed tomography (MVCT) during image-guided intensity-modulated radiation therapy (IMRT) for crani spinal irradiation (CSI) on helical tomotherapy (HT). *Technol. Cancer Res. Treat.* **14**(1), 29–36 (2015)
2. Ruchala, K.J., Olivera, G.H., Schloesser, E.A., Mackie, T.R.: Megavoltage CT on a tomotherapy system. *Phys. Med. Biol.* **44**(10), 2597–2621 (1999)
3. Yu, Z.H., Kudchadker, R., Dong, L., Zhang, Y., Court, L.E., Mourtada, F., Yock, A., Tucker, S.L., Yang, J.: Learning anatomy changes from patient populations to create artificial CT images for voxel-level validation of deformable image registration. *J. Appl. Clin. Med. Phys.* **17**(1), 246–258 (2016)
4. De Los Santos, J., Popple, R., Agazaryan, N., Bayouth, J.E., Bissonnette, J.P., Bucci, M.K., Dieterich, S., Dong, L., Forster, K.M., Indelicato, D., Langen, K., Lehmann, J., Mayr, N., Parsai, I., Salter, W., Tomblyn, M., Yuh, W.T., Chetty, I.J.: Image guided radiation therapy (IGRT) technologies for radiation therapy localization and delivery. *Int. J. Radiat. Oncol. Biol. Phys.* **87**(1), 33–45 (2013)

5. Jeraj, R., Mackie, T.R., Balog, J., Olivera, G., Pearson, D., Kapatoes, J., Ruchala, K., Reckwerdt, P.: Radiation characteristics of helical tomotherapy. *Med. Phys.* **31**(2), 396–404 (2004)
6. Shah, A.P., Langen, K.M., Ruchala, K.J., Cox, A., Kupelian, P.A., Meeks, S.L.: Patient dose from megavoltage computed tomography imaging. *Int. J. Radiat. Oncol. Biol. Phys.* **70**(5), 1579–1587 (2008)
7. Gao, H., Qi, X.S., Gao, Y., Low, D.A.: Megavoltage CT imaging quality improvement on TomoTherapy via tensor framelet. *Med. Phys.* **40**(8), 081919 (2013)
8. Lu, W., Olivera, G.H., Chen, Q., Ruchala, K.J., Haimerl, J., Meeks, S.L., Langen, K.M., Kupelian, P.A.: Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. *Phys. Med. Biol.* **51**(17), 4357–4374 (2006)
9. Sheng, K., Gou, S., Wu, J., Qi, S.X.: Denoised and texture enhanced MVCT to improve soft tissue conspicuity. *Med. Phys.* **41**(10), 101916 (2014)
10. Yaru, L., Chenxi, Y., Jian, Z., Haining, Y., Yang, C., Yong, Y., Baosheng, L., Jiwen, D.: A Megavoltage CT image enhancement method for image-guided and adaptive helical TomoTherapy. *Front. Oncol.* **9**, 362 (2019)
11. Phillip, I., Junyan, Z., Tinghui, Z., Alexei A. Efros.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
12. Jie, C., Zibo, M., Chiuman, H.: Residual channel attention generative adversarial network for image super-resolution and noise reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 454–455 (2020)
13. Rui, L., Yixiao, G., Chinglam, C., Xiaogang, W., Hongsheng, L.: DivCo: diverse conditional image synthesis via contrastive generative adversarial network, arXiv preprint [arXiv:2103.07893](https://arxiv.org/abs/2103.07893) (2021)
14. Nripendra Kumar, S., Khalid, R.: Medical image generation using generative adversarial networks, arXiv preprint [arXiv:2005.10687](https://arxiv.org/abs/2005.10687) (2020)
15. Xin, Y., Ekta, W., Paul, B.: Generative adversarial network in medical imaging: a review. *Med. Image Anal.* **58**, 101552 (2019)
16. Junyan, Z., Taesung, P., Phillip, I., Alexei A. Efros.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
17. Vinas, L., Scholey, J., Descovich, M., Kearney, V., Sudhyadhom, A.: Improved contrast and noise of megavoltage computed tomography (MVCT) through cycle-consistent generative machine learning. *Med. Phys.* **48**(2), 676–690 (2021)
18. Xiaojiao, M., Chunhua, S., Yubin, Y.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. arXiv preprint [arXiv:1603.09056](https://arxiv.org/abs/1603.09056) (2016)
19. Hu, C., Yi, Z., Mannudeep, K., Kalra, M., Feng, L., Yang, C., Peix, L., Jiliu, Z., Ge, W.: Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans. Med. Imag.* **36**(12), 2524–2535 (2017)
20. Chao, D., Chenchange, L., Kaiming, H., Xiaou, T.: Image super-resolution using deep convolutional networks. In: IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
21. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29**(1), 196–205 (2010)

# Research and Application of Railway Turnout Gap Detection Based on Improved Canny Algorithm



Xinpeng Liu , Runyuan Sun , and Zhifeng Liang

**Abstract** Turnout gap monitoring is research emphasis and hotspots of railway safety management. In order to solve the automation requirements of the turnout gap monitoring system, this paper proposes an automatic turnout gap detection algorithm based on image processing technology. After image preprocessing, an edge detection algorithm based on the Canny arithmetic operator is used to convert the image into an edge binary image, then look for all possible gaps. Then after screening, the real gap is preserved and marked. To solve the problem of low accuracy with complex environment, an improved Canny edge detection algorithm is proposed. The detection experimental results on images with different noises show that this method is better than traditional edge detection algorithm and get a high robustness in noisy environment. Meanwhile, this method can successfully eliminate the stripe noise.

**Keywords** Canny operator · Edge detection · Turnout gap monitoring

## 1 Introduction

Turnouts are one of the important equipment at the railway station's electrical service site, the reliability and safety of their working conditions are directly related to the safe operation of railway transportation [1]. The switch machine is the core equipment for railway turnouts and can be used to realize the switch and lock function of the turnout, the working state of the switch machine plays a decisive role in the speed of the train and the safety of operation [2, 3]. In the monitoring of the switch machine, the gap of the switch machine is an important parameter of the working

---

Supported by: Science and Technology Program of University of Jinan (XKY1930).

---

X. Liu · R. Sun ()

School of Information Science and Engineering, University of Jinan, Jinan 250022, People's Republic of China

e-mail: [sunry@ujn.edu.cn](mailto:sunry@ujn.edu.cn)

X. Liu · R. Sun · Z. Liang

Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, People's Republic of China

state of the switch machine. It reflects the degree of adhesion between the basic rail and the switch rail after the switch is switched between the positioning and the reverse position [4]. To measure the gap in turnout, researchers have used a variety of methods. Some traditional methods [5–7] use electrical measurement methods, which are not accurate enough and easily affect other equipment. In recent years, with the development of computer technology, many researchers have tried to use image.

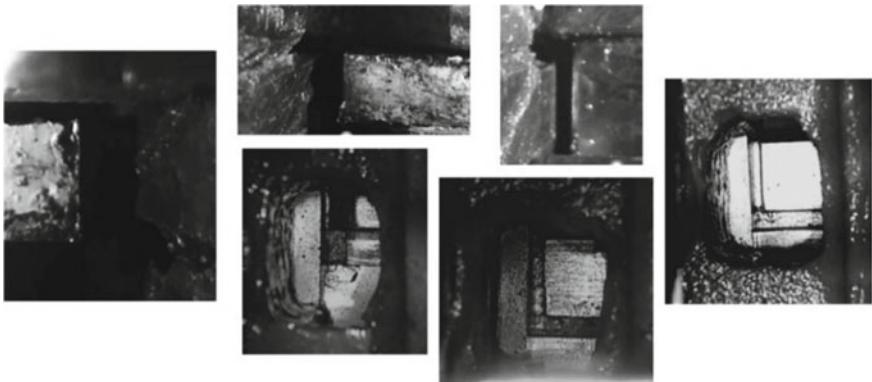
processing technology, deep learning, and other methods to realize the measurement of the gap in turnout [8–10]. Compare with the traditional method, the degree of automation is higher.

However, the working environment of the switch machine is relatively harsh, which will have a greater impact on the accuracy of most existing algorithms. This paper proposes an automatic switch gap detection algorithm based on image processing technology. First, after preprocessing the image, use the edge detection algorithm to detect the edge of the image, then look for possible gaps in the image, and finally filter the results. In this paper, the widely used Canny operator edge detection algorithm is optimized for the shortcomings of noise sensitivity. In the preprocessing stage, more operations are performed on the image to eliminate noise, and the method of calculating weights through multiple detections is used to find potential gaps, Improve the accuracy of detection. Experiments have proved that this method has high accuracy and reliability, and does not need to install identification components on the gap. Compared with the original Canny operator edge detection algorithm, its anti-noise ability is stronger.

## 2 Turnout Gap Detection Method and System Principle

The turnout gap is the distance between the indication rod and the detection column inside the switch machine. The gap of the turnout characterizes the degree of close contact between the turnout and the basic rail. Therefore, measuring the gap can check whether the switch machines can change in place each time the track changes.

To measure the turnout gap, one of the methods usually used now is to install a camera device inside the switch machine to continuously take pictures of the part of the indication rod detection block. As shown in Fig. 1, since the internal structure of different types of switch machines is very different, the camera installation angle and the distance from the gap are also different. Therefore, it is usually to manually determine the detection area and calibrate the baseline before the measurement., to speed up the detection speed and improve accuracy.



**Fig. 1** Internal gaps of different models of switch machines

### 3 Automatic Detection Algorithm for Turnout Gap

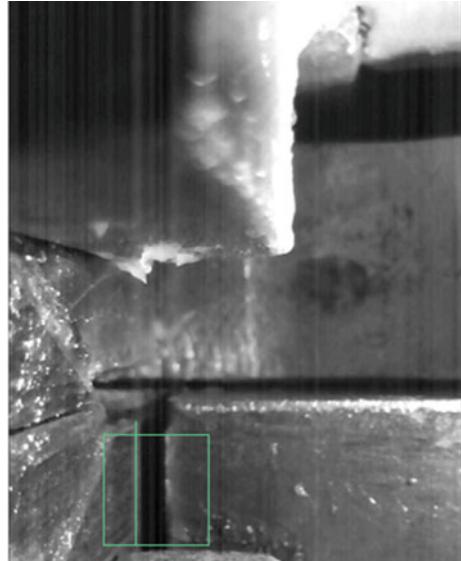
#### 3.1 Parameter Setting

The parameter setting is mainly to manually delimit the detection area and the baseline. One-time parameter setting can be effective for a group of continuously gap pictures of turnouts for subsequent gap recognition. The baseline is the edge on the side of the bar in the gap, generally, will not change position in the image. The delineation of the detection area can be freely delineated manually, or the entire gap can be drawn when the baseline is drawn, and then the one where the baseline is located is directly taken as the detection area. This method is simpler and only requires manual delineation. The baseline is sufficient, as shown in Fig. 2.

#### 3.2 Edge Detection

Before looking for the gap, the image needs to be converted into a meaningful binary image, which is generally realized by using an edge detection algorithm. The Canny operator is relatively effective when extracting image edges from noisy images [11], so the algorithm uses the Canny operator to extract the edges of images. The edge detection algorithm based on the Canny operator mainly has four steps, namely image denoising, gradient calculation, non-maximum suppression, double threshold detection and edge connection.

**Image denoising.** In order to remove the small noise in the image and smooth the image, use Gaussian filter to filter the image. The formula of the Gaussian filter convolution kernel is:

**Fig. 2** Setting of parameters

$$h(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (1)$$

Let  $f(x, y)$  be the original image, then the filtered image  $g(x, y)$  is:

$$g(x, y) = f(x, y) * h(x, y, \sigma) \quad (2)$$

\* is the convolution operation, generally, take  $\sigma = 2$ .

Cameras often leave striped white noise on the image due to the refreshing of the image when shooting. In order to eliminate the influence of stripe noise, it is necessary to use the dilation and erosion of morphological operations. For a given image  $g(x, y)$  and structural element S, expanding each point  $(x, y)$  in the image into an area the size of structural element S is called a dilation operation on image  $g(x, y)$ , on the contrary, each area congruent with the structural element S is contracted into a point, which is called the erosion operation on the image  $g(x, y)$ . Performing multiple erosions on the image and then performing the same number of dilation operations can remove small blocks of noise in the image, can effectively remove stripe noise, called open operation. The effect depends on the number of operations, e. After investigation and some experiments, it is found that  $e = 1$  has a better denoising effect on most pictures, and can keep the error caused by image destruction at a small level.

**Gradient calculation.** The edge is often the point where the gray level of the image changes greatly, so the edge detection algorithm based on the Canny operator needs to calculate the gradient of each point  $(x, y)$ . The calculation method of using the Sobel operator to calculate the gradient amplitude  $M(x, y)$  and the angle  $\theta_M$  is:

$$\begin{aligned} M(x, y) &= |d_x(x, y)| + |d_y(x, y)| \\ &= |f(x, y) * Sobel_x| + |f(x, y) * Sobel_y| \end{aligned} \quad (3)$$

$$\theta_M = \arctan\left(\frac{d_y}{d_x}\right) \quad (4)$$

Among them,  $d_x$  and  $d_y$  are the gradient magnitude in the horizontal and vertical directions, respectively,  $Sobel_x = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$ ,  $Sobel_y = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$ .

**Non-maximum suppression.** Only perform edge extraction based on the gradient, the edge will be very blurred. The function of non-maximum suppression is to suppress gradients other than the local maximum to zero. For each pixel, compare its gradient magnitude with two adjacent pixels in the positive and negative gradient direction. If the gradient magnitude is the largest, it will remain as an edge point, otherwise the pixel will be suppressed. Usually, the pixels to be compared are obtained by calculating linear interpolation. As shown in Fig. 3, each pixel has 8 adjacent pixels, and their pixel values are E, NE, N, NW, W, SW, S, SE, and the numbers represent the angle area. The gradient direction of the pixel point P is  $\theta$ .

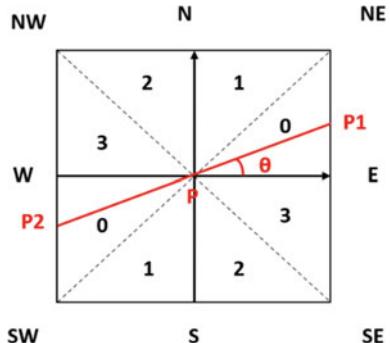
Taking the gradient direction in the 0 areas as an example, let  $d_x$  and  $d_y$  be the gradient magnitudes in the horizontal and vertical directions respectively, then the calculation method of linear interpolation  $d_{p1}$  and  $d_{p2}$  is:

$$\tan \theta = \frac{d_y}{d_x} \quad (5)$$

$$d_{p1} = 1 - \tan \theta \cdot E + \tan \theta \cdot NE \quad (6)$$

$$d_{p2} = 1 - \tan \theta \cdot W + \tan \theta \cdot SW \quad (7)$$

**Fig. 3** Calculating linear interpolation



If and only if the gradient magnitude of the current pixel is greater than  $d_{p1}$  and  $d_{p2}$ , the pixel is considered to be an edge, otherwise the pixel is suppressed.

**Double threshold detection and edge connection.** The edge detection algorithm based on the Canny operator generally needs to input two threshold parameters, one high and one low, to adjust the effect of edge detection. For all possible edge pixels, compare their gradient magnitudes with the high and low thresholds. The pixels greater than the high threshold are marked as strong edge pixels and determined as edges; the points less than the low threshold will be suppressed; for the points less than the high threshold and greater than the low threshold, marked as weak edge pixels, using the 8-connected area method, when there are edge pixels in the surrounding 8 pixels, the pixels will be marked as strong edge pixels, otherwise suppressed.

### 3.3 Gap Detection

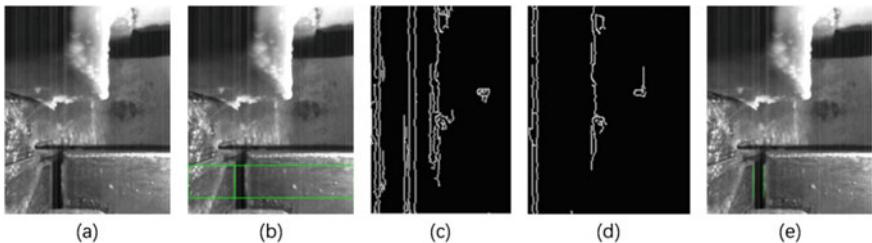
The image after edge detection is a binary image, so the subsequent operation on the binary image is used to find the gap. There are two steps to find the gap. First traverse the image from the calibration baseline to find possible gap lines, and then, filter out false gaps caused by interference through certain conditions.

**Looking for gaps.** To find the gap, set the baseline as a line segment from point  $(x_b, y_{b1})$  to point  $(x_b, y_{b2})$ , the edge of the image in the direction of the gap is  $x_{end}$ , and the search algorithm flow is, set  $f(x, y)$  as a binary value For the image, set  $x_{start} = x_b$  so that all pixels from  $y_{b1}$  to  $y_{b2}$  traverse the  $x_{end}$  direction from  $x_b$ , and mark the x-axis coordinates of the first edge point encountered. After completing a round, take the median of all coordinates as the result of this round of finding gaps, and then use the result coordinates this time as the new  $x_{start}$ , and repeat the above steps until all possible gaps are found.

**Calibration gap.** After obtaining the gap set, there's a need to calculate their weights to select the real gaps. The specific method is to find the boundary to left and right for each possible gap and calculate the width. Then calculate the length of the gap. First, filter out those with too short length and incorrect shape, and then calculate the weight = length \* pixel value from the left boundary to the right boundary, and finally select the one with the largest weight as the true gap.

## 4 Experimental Verification

In order to verify the effectiveness of the algorithm, the experiment selects the real photos of the switch machine at Hepu Station for verification. As shown in Fig. 4, a picture of the ZY6 switch machine is selected for the algorithm experiment.



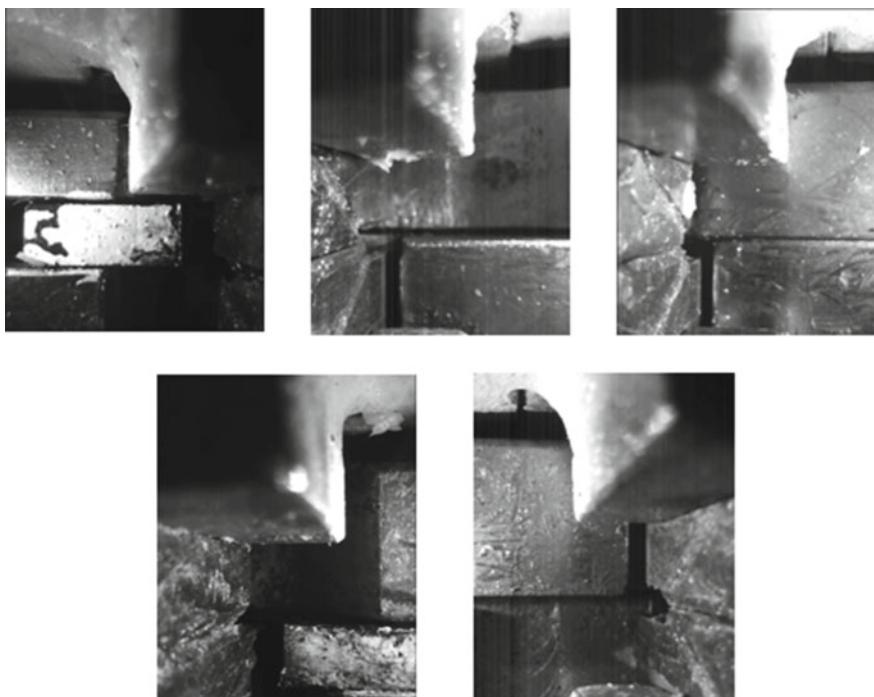
**Fig. 4** Algorithm experiment

It can be seen from the figure that the original image (Fig. 4a) is dimly lit and there are serious noise and fringe noise interference in the picture, and the image quality is low, which is convenient for detecting the robustness of the algorithm. As shown in Fig. 4b, set the parameters for the original image, and the effect after native edge detection is shown in Fig. 4c. It can be seen that there are still many unremoved noises in the image, which may cause errors in the algorithm. After using the improved edge detection algorithm, the effect is shown in Fig. 4d. It can be seen that compared with Fig. 4c, most of the noise is eliminated. The result of gap detection is shown in Fig. 4e. The comparison error is within the acceptable range. The algorithm is robust.

To test the effectiveness of the algorithm in different environments and different models, the experiment selected 7 data sets from two switch machines for testing. The images taken in the switch machine are shown in Fig. 5, the data set information is shown in Table 1.

The algorithm improves the edge detection algorithm based on the Canny operator. To test the improvement effect of the algorithm, the experiment uses the original Canny operator edge detection algorithm and the improved edge detection algorithm to compare the accuracy and average error of the test. The results are as follows (Figs. 6 and 7; Tables 2 and 3).

It can be seen that the overall accuracy of the algorithm is relatively high, which proves that the design of the algorithm for finding the gap is successful. In contrast to the edge detection algorithm, the accuracy of the improved algorithm is significantly higher than that of the native Canny operator. Due to the error of manual labeling, the value of the average error can only be used as a reference. In comparison, it can be found that the average error of the improved algorithm is lower than that of the original algorithm in most cases, but some situations are slightly higher than the original algorithm. The reason may be that the improved algorithm has caused some damage to the image quality while removing the noise, and will produce subtle errors. However, in general, the improved edge detection algorithm is more effective than the original Canny operator edge detection algorithm and is more suitable for complex and changeable actual production environments.



**Fig. 5** Switch machines in different scenarios

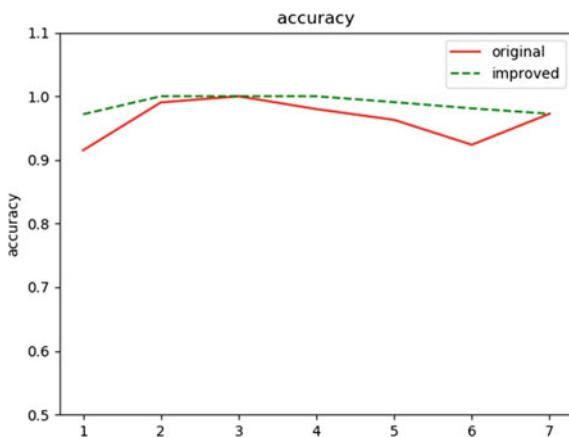
**Table 1** Data set information

Number	Number of pictures
1	106
2	103
3	106
4	99
5	108
6	105
7	109

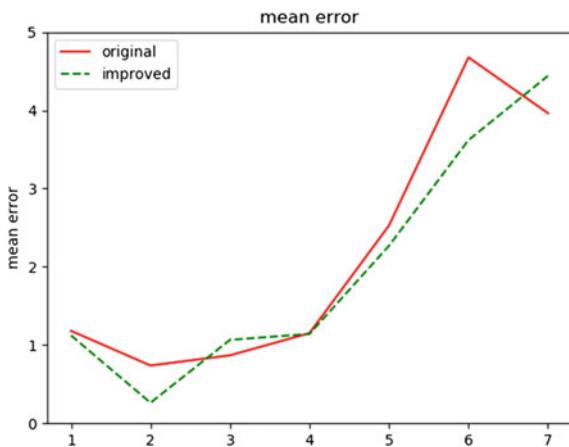
## 5 Conclusions

An important topic of railway turnout safety monitoring is machine gap monitoring. To measure the gaps, researchers have used many methods. One of the widely used methods is to use camera equipment to shoot the gaps and then use edge detection algorithms. Perform gap recognition, but due to the complex working environment of the switch machine, the traditional edge detection algorithm has been greatly challenged. In this paper, the original Canny edge detection algorithm is improved, the

**Fig. 6** Accuracy rate comparison



**Fig. 7** Mean error comparison



**Table 2** Accuracy data

Number	Origin	Improved
1	0.9151	0.9717
2	0.9903	1
3	1	1
4	0.9798	1
5	0.963	0.9907
6	0.9238	0.981
7	0.9725	0.9725

**Table 3** Mean error data

Number	Origin	Improved
1	1.4811	1.1226
2	0.7379	0.2621
3	0.8679	1.066
4	1.1515	1.1414
5	2.5278	2.2685
6	4.6762	3.6190
7	3.9633	4.4404

image preprocessing stage and gap detection stage is improved and some algorithms are designed, and use the point machine gap shooting data with noise in different environments for algorithm verification. Experiments show that this method has high overall accuracy and reliability, and can still achieve better results on image data with very low image quality, but there is still room for improvement. The algorithm can run well in different environments and can meet the needs of turnout gap monitoring in actual production. The application of the algorithm can contribute to the maintenance of the safety of the railway system.

## References

- Dindar, S., Kaewunruen, S., An, M.: Identification of appropriate risk analysis techniques for railway turnout systems. *J. Risk Res.* **21**(8), 974–995 (2018)
- Hamadache, M., Dutta, S., Olaby, O., et al.: On the fault detection and diagnosis of railway switch and crossing systems: an overview. *Appl. Sci.* **9**(23), 5129 (2019)
- Izadi, I., Shah, S.L., Shook, D.S., et al.: An introduction to alarm analysis and design. *IFAC Proc.* **42**(8), 645–650 (2009)
- de Aguiar, E.P., Fernando, M.A., Amaral, R.P.F., et al.: EANN 2014: a fuzzy logic system trained by conjugate gradient methods for fault classification in a switch machine. *Neural Comput. Appl.* **27**(5), 1175–1189 (2016)
- Asada, T., Roberts, C., Koseki, T.: An algorithm for improved performance of railway condition monitoring equipment: Alternating-current switch machine case study. *Transport. Res. Part C: Emerg. Technol.* **30**, 81–92 (2013)
- Bacchelli, S., Papi, S.: Filtered wavelet thresholding methods. *J. Comput. Appl. Math.* **164**, 39–52 (2004)
- Igarashi, Y., Siomi, S.: Development of monitoring system for electric switch machine. *Quart. Rep. RTRI* **47**(2), 78–82 (2006)
- Tao, T., Dong, D., Huang, S., et al.: Gap detection of switch machines in complex environment based on object detection and image processing. *J. Transport. Eng. Part A: Syst.* **146**(8), 04020083 (2020)
- Wang, C., Liu, Q., Wang, W.: Design and implementation of monitoring controller for switch machine gap based on image processing. In: 2019 Chinese Control Conference (CCC), pp. 6610–6614. IEEE (2019)

10. Zhu, C., Xu, Y.: Research on intelligent monitoring of switch machine based on image processing technology. In: International Conference on Transportation and Development 2020, pp. 161–167. American Society of Civil Engineers, Reston, VA (2020)
11. Joshi, M., Vyas, A.: Comparison of Canny edge detector with Sobel and Prewitt edge detector using different image formats. Int. J. Eng. Technol. 133–137 (2020)

# 3D Vision Transformer for Postoperative Recurrence Risk Prediction of Liver Cancer



Fan Li , Xueying Zhou , Xizhan Gao , Hui Zhao , and Sijie Niu

**Abstract** Hepatocellular carcinoma (HCC) is a kind of malignant tumor with a high fatality rate, and it has a serious impact on the patient's normal life. Although the diagnostic scheme for liver cancer has been gradually improved in recent years, the prognosis of patients with liver cancer is still not very ideal. Due to the high recurrence rate and poor prognosis after surgery, intrahepatic metastasis plays an important role in the survival cycle of patients with liver cancer, which is also a significant guiding factor in the selection of preoperative treatment and the planning of postoperative follow-up. In this paper, a classification method based on 3D Vision Transformer is proposed to determine whether intrahepatic metastasis occurs after the operation. Specifically, in this study, 161 patients who were diagnosed as HCC by puncture pathology or clinical diagnosis and received Transcatheter Arterial Chemoembolization (TACE) treatment were selected as the study subjects, and the patients were divided into non-metastasis group (79 cases) and intrahepatic metastasis group (82 cases). By comparing 3D Vision Transformer with the 3D forms of multiple deep learning classification models, the predictive value of 3D Vision Transformer in predicting intrahepatic metastasis after transarterial chemoembolization in HCC patients was verified.

**Keywords** 3D vision transformer · Deep neural network · Classification · Hepatocellular carcinoma

## 1 Introduction

Liver cancer is one of the most common cancers and major cause of cancer deaths in China, which accounts for over 50% of new cases and deaths worldwide [1]. Primary liver cancer is divided into two major types—hepatocellular carcinoma

---

F. Li · X. Zhou · X. Gao · H. Zhao · S. Niu (✉)

School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

(HCC) derived from the hepatocytes and cholangiocarcinoma from the intrahepatic bile ducts [2]. HCC accounts for 90% of primary liver cancer, which is the fourth most common cause of cancer-related deaths worldwide [3]. How to evaluate the effect of treatment, improve the level of prognosis, improve the quality of life of patients and extend the life cycle of patients are urgent problems to be solved. Therefore, it is of great importance to carry out the research on the diagnosis and prognosis of liver cancer in China. It can not only help the patients with liver cancer to complete the earlier diagnosis of liver cancer, but also help doctors to make an individualized treatment plan for them.

Hamamoto et al. [4] conducted a study on machine learning prediction on HCC partial hepatectomy to determine the value machine learning for prediction. Subsequently, researchers have done a lot of research on the quality of life, disease-free survival, and overall survival in patients with HCC after partial hepatectomy. The result shows that machine learning can accurately predict, which is better than traditional analytical methods [5–8]. Deep Learning is a new direction of the development of machine learning, which is influencing every field of people's life. Liu et al. [9], based on the deep learning, enhanced ultrasound method to accurately predict the treatment effect of HCC patients after TACE. Peng et al. [10] used convolution neural network (CNN) to effectively predict the treatment effect of HCC after TACE through CT imaging. Therefore, deep learning models can well predict the effect of TACE treatment and provide the basis for clinicians to select surgical cases more efficiently.

Most existing deep learning-based methods for classification are based on CNN, such as ResNet [11], DenseNet [12] and so on. But despite the excellent representational capabilities of the CNN-based methods, due to the inherent limitations of convolution operations, they often show limitations in establishing remote dependencies, this is especially true for textures, shapes, and sizes that can show significant differences between patients. Transformer is known for modeling remote dependencies in data. Recently, because of the use of the Transformer structure, there are a large number of computer vision tasks results have achieved outstanding performance, such as Vision Transformer (ViT) [13], Swin Transformer [14]. Therefore, Transformer becomes the first-choice scheme in the field of computer vision and this structure is opening another door to an image classification algorithm. Motivated by this, we believe the application of Transformer to the field of postoperative recurrence risk prediction of liver cancer is worth exploring. Based on the above discussion, we propose a medical 3D image classification model based on ViT and we call it 3D Vision Transformer (3D ViT). Specifically, our proposed method has three main contributions:

- (1) Since the ViT is a model that only receive two-dimensional data, the network we proposed in this paper can take 3D volumes as input.
- (2) Vision Transformer as a current mainstream classification method, we applied it for the first time to predict intrahepatic metastasis in HCC patients after TACE treatment.

- (3) Compared with the previous classification models, our proposed model achieves excellent results on our magnetic resonance imaging (MRI) dataset for HCC patients treated with TACE.

## 2 Methodology

### 2.1 Data Pre-processing

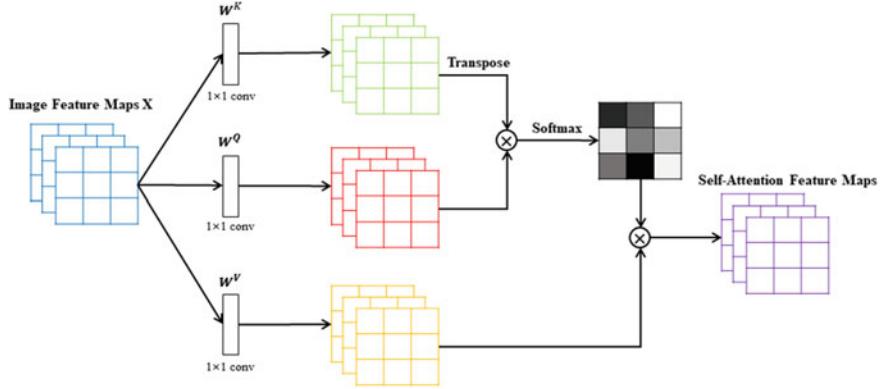
The images were collected using 3 T magnetic field intensity, and different patients were scanned with different MR image scanners (i.e., Siemens Medical Systems and Philips Medical Systems). In this case, due to the quite difference in shape and appearance of MR images, we resampled the volume spacing of our data to 1 mm × 1 mm × 3 mm. Then these MR images are resized to 128 × 128 × 96. Meanwhile, we also use some data augmentation method for the input data, like random crop, random rotate and random horizontal flip.

### 2.2 Self-attention and Multi-head Attention

**Self-Attention.** For the common attention mechanisms in computer vision tasks, such as SENet [15] and CBAM [16], attention distribution is usually calculated in channel or spatial of the image. The self-attention mechanism in image processing is an idea borrowed from natural language processing (NLP) tasks, and it is also an indispensable part of Transformer. Therefore, the names of Query, Key and Value are retained. Self-attention can obtain the spatial dependency of any two positions in the feature map and acquire the long-distance context information. Using  $X \in R^{n \times d}$  to represent a sequence of  $n$  entities  $(x_1, x_2, \dots, x_n)$ , where  $d$  represents the embedding dimension of each entity. The goal of self-attention is to encode each entity with global context information to obtain the relationships between all  $n$  entities. The representation of this relationship can be realized by three learnable weight matrices: Queries, Keys and Values. The mathematical symbols can be described as  $W^Q \in R^{d \times d_k}$ ,  $W^K \in R^{d \times d_k}$ ,  $W^V \in R^{d \times d_v}$  respectively.

Firstly, the input sequence  $X$  is projected onto these weight matrices to get  $Q = XW^Q$ ,  $K = XW^K$ ,  $V = XW^V$ . Because it is a single head self-attention, that is  $d_q = d_k = d_v = d$ . Then the output of the self-attention layer is formulated in Eq. 1:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$



**Fig. 1** An example of self-attention block used in the computer vision tasks

As shown in Fig. 1, we calculate the dot-product between the query and all keys, then normalized by softmax operator function to get the attention score, the normalized weights are weighted sum with the corresponding values to obtain the final output of self-attention layer.

**Multi-head Attention.** In order to compress the complex relations between different elements in a sequence, multi-head attention includes several self-attention modules. Multi-head attention is similar to the idea of using multiple convolution kernel in the same layer in CNN. Each attention head only focuses on one subspace of the final output sequence and independent with each other. The purpose of this structure is hope to extract more abundant potential features, as shown in Eq. 2.

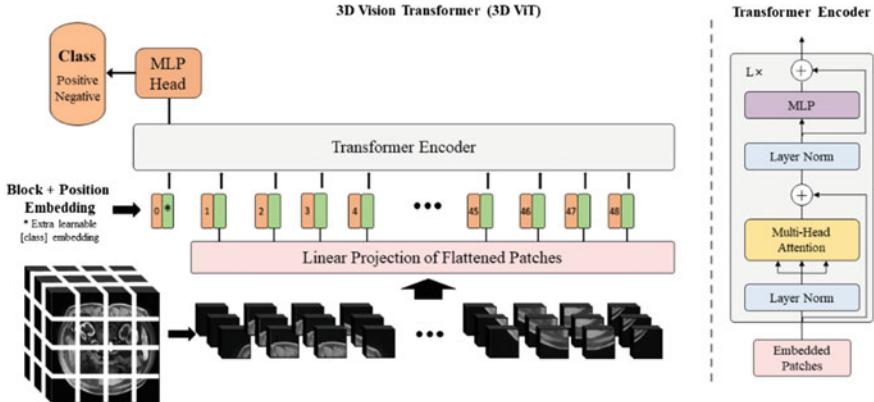
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O \quad (2)$$

$$\text{where, } \text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad i = 1 \dots h \quad (3)$$

where  $W_i^Q, W_i^K \in R^{d \times d_k}$ ,  $W_i^V \in R^{d \times d_v}$ ,  $\text{Concat}(\cdot) \in R^{n \times h \cdot d_v}$ ,  $W^O \in R^{h \cdot d_v \times d}$ . In self-attention, it can be seen as only one head, while multi-head attention is to split the generated  $Q$ ,  $K$ ,  $V$  matrices into multiple small matrices to jointly information of common concern from different representation subspaces at different positions.

### 2.3 3D Vision Transformer

Transformer is widely used in natural language processing field, such as machine translation, question answering system. Vision Transformer is an architecture based on Transformer, which is the first work to using Transformer completely replace standard convolution in deep neural networks on large scale computer vision datasets.



**Fig. 2** An overview of 3D Vision Transformer (the left) and the internal structure of Transformer encoder (the right). the sample MR image size is  $128 \times 128 \times 96$ , we split these images into a fixed-size (i.e.,  $32 \times 32 \times 32$ ) block, linearly embed each of them and attached position embeddings, then feed these vectors to a standard Transformer encoder

The main goal is to generalize them to image formats without integrating any specific data architecture. They applied the original Transformer model on a sequence of image patches. Meanwhile, different from the CNN architecture with local receptive field filters commonly used in computer vision field, the self-attention mechanism used by the Vision Transformer allows it to focus on information of the whole image.

On the basis of the original Vision Transformer, we directly constructed the 3D Vision Transformer (3D ViT) by changing the size of the input images, the modified network structure shown in Fig. 2. To handle 3D images, we divide the three-dimensional image into several fixed-size image blocks and flatten each image block into vector form, then the linear layer is used to project to a block embedding for these vectorized image blocks. Meanwhile, location information is attached with it by position embedding method. A token stands for classification is appended at the beginning of the input sequence to Transformer encoder.

Let's denote a preprocessed MR image by  $X \in \mathbb{R}^{D \times H \times W \times C}$ , where  $(D, H, W)$  is the resolution of the original image,  $C$  is the number of channels. Suppose that each fixed-size block is  $P \times P \times P \times C$  in size, so each small piece can be defined as  $X_p \in \mathbb{R}^{N \times (P^3 \cdot C)}$ , where  $(P, P, P)$  is the resolution of each image block, in this paper, we set  $P$  is 16.  $N = DHW/P^3$  refers to the total number of blocks, which also can be taken as the effective length of the input sequence for Transformer encoder. Then we flatten the image blocks and map to  $D$  dimensions by a trainable linear projection. We call the output of this projection as patch embeddings. In order to achieve classification, we add a learnable classification token  $X_{class}$  at the beginning of the sequence, as shown in Eq. 4. At the same time, according to the convention of Transformer's positional encoding, 3D ViT also introduces position embeddings, which is a learnable vector attached to patch embedding to retain positional information, it denoted by the symbol  $E_{pos}$ .

The subsequent processing is the same as in Transformer encoder. It consists of multiple multi-head attention blocks (MSA), MLP modules, and LayerNorm (LN) which is added before every block, also we use residual connections after every block (Eqs. 5 and 6). In the MLP module, which contains two fully connected layers with a GELU activation function behind each layer.

Finally, the first position of the Transformer encoder's output  $z_L^0$  is used as the global image representation for image classification task and then sent to MLP to output classification result.

$$z_0 = [X_{class}; X_p^1 E; X_p^2; \dots; X_p^N E] + E_{pos}, \quad E \in R^{(P^3 \cdot C) \times D}, E_{pos} \in R^{(N+1) \times D} \quad (4)$$

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1 \dots L \quad (5)$$

$$z_l = MLP(LN(z'_l)) + z'_l, \quad l = 1 \dots L \quad (6)$$

In this paper, for our model, the input size of MR images is  $128 \times 128 \times 96$  and we set  $P$  to 16. The number of output categories is 2 (transferred or not transferred). Binary cross entropy loss is used to our task. The loss function is formulated in Eq. 7:

$$L_{BCE} = - \sum_{i=1}^2 t_i \log(p_i) = -[t \log(p) + (1-t) \log(1-p)] \quad (7)$$

where  $t_i$  is the truth value taking a value 0 or 1 and  $p_i$  is the softmax probability for the  $i$ th class.

### 3 Experimental Results

To evaluate the proposed method in the postoperative liver cancer metastasis prediction task, we apply the 3D ViT on own dataset and compared it with the multiple mainstream models of vision tasks, which will be introduced in later subsection.

#### 3.1 Dataset and Metrics

Patients with HCC who underwent TACE treatment between January 2015 and December 2019 were retrospectively collected. All patients underwent upper abdominal plain and enhanced CT scan before TACE treatment, but only in the initial TACE patients, and these patients underwent TACE treatment at least twice. Inclusion criteria: (1) preoperative biopsy or clinical diagnosis of HCC; (2) transcatheter

arterial chemoembolization; (3) no more than 3 multiple tumors with a maximum lesion diameter of 3 cm; (4) no single lesion with a maximum lesion diameter of 5 cm; (5) survival data were obtained by follow-up. Exclusion criteria: (1) diffuse or giant tumor; (2) other liver malignancies confirmed by pathology; (3) radiation therapy, chemotherapy and liver transplantation before operation or during follow-up; (4) before the third TACE treatment in patients with portal vein tumor thrombus; (5) MRI image motion artifacts, affect the judgment.

A total of 161 eligible patients, 136 males and 25 females, aged 27–80, were included in the analysis. The final follow-up time of the case is September 2020. The patient's case review and telephone follow-up through the hospital medical record inquiry system fully understand the survival and recurrence of the patient within 3 years after discharge from hospital, the patients were divided into non-metastasis group (79 cases) and intrahepatic metastasis group (82 cases).

Images were collected using a superconducting magnetic resonance scanner, and the upper abdomen was scanned using a 16-channel body coil, including plain and enhanced MRI scans. Each patient had a T1-weighted MR image and its corresponding label given by the specialist to determine if the patient had intrahepatic metastases. The images were acquired with a 3 T magnetic field intensity, but different patients were scanned using different MR image scanner. (i.e., Siemens Medical Systems and Philips Medical Systems). In this case, the image is quite different in shape and appearance, so we resample the collective pixel spacing of our data to  $1 \text{ mm} \times 1 \text{ mm} \times 3 \text{ mm}$ , then resize the image to  $128 \times 128 \times 96$ .

Furthermore, accuracy and recall rate as our evaluation criteria are used to assessment the performance of our model. Accuracy refers to the proportion of correctly of classified samples to the total number of samples, as shown in Eq. 8.

$$\text{Accuracy} = \frac{n_{\text{correct}}}{n_{\text{total}}} \quad (8)$$

where  $n_{\text{correct}}$  is the number of samples correctly classified,  $n_{\text{total}}$  is the total number of samples. Recall rate refers to the proportion of correctly classified positive samples to the number of truly positive samples, as shown in Eq. 9.

$$\text{Recall} = \frac{n_{TP}}{n_{TP+FP}} \quad (9)$$

where  $n_{TP}$  is the number of positive sample correctly classified,  $n_{TP+FP}$ . indicates the total number of positive samples.

### 3.2 Implementation Details

Pytorch is adopted to implement 3D Vision Transformer shown in Fig. 2. The model training and testing hardware platform is Ubuntu 18.04.5 LTS, and a Titan T4 GPU

**Table 1** Details of 3D vision transformer model variants

Model	Blocks	Hidden size	MLP size	Heads
3D ViT-base	6	512	2048	12
3D ViT-large	12	1024	4096	16

**Table 2** The results of different methods on our dataset

Model	Accuracy (%)	Recall (%)	Params
3D ResNet50	$56.45 \pm 3.81$	$55.77 \pm 4.42$	46 M
3D ResNet101	$60.82 \pm 3.77$	$59.13 \pm 4.10$	86 M
3D SeResNext50	$63.08 \pm 6.19$	$64.03 \pm 6.32$	28 M
3D SeResNext101	$64.64 \pm 7.18$	$64.19 \pm 6.98$	51 M
3D ViT-Base (Ours)	$63.44 \pm 8.42$	$64.50 \pm 8.54$	24 M
3D ViT-Large (Ours)	<b><math>66.17 \pm 8.06</math></b>	<b><math>65.47 \pm 8.32</math></b>	156 M

is utilized to train the network. In the training phase, a fivefold cross validation procedure is performed on the 161 patients, for each fold, the patients were divided into 129 training patients and 32 validation patients. We adopt SGD algorithm with momentum to optimize the network. For this classification task, we train 200 epochs on training dataset in every fold, the size of input is  $128 \times 128 \times 96$  and the batch size is set to be 8. The learning rate for the network is initialized to  $1e-4$  and cosine learning rate scheduler is used to adjust the learning rate.

### 3.3 Comparison with Existing Methods

As shown in Table 1, referred to ViT model variant, we also launched two 3D ViT model variants through experiments to adjust the setting of the super parameters.

As shown in Table 2, compared with the models that have performed well in ImageNet. We report mean and standard deviation of accuracy and recall rate, our method has achieved relatively better results. Specifically, by comparing 3D ViT-Base model and 3D ViT-Large model, it can be found that adjusting the super parameters configuration of 3D ViT model can effectively improve the pre-diction results. Meanwhile, 3D ViT-Large model is better than other model we compared in accuracy and recall rate, but its parameters are also quite large.

## 4 Conclusion and Future Work

In this paper, we explore the predictive of 3D Vision Transformer in HCC patients with intrahepatic metastasis after transcatheter arterial chemoembolization. The

experimental results show that 3D Vision Transformer is competitive with some recent deep learning classification methods. In the meantime, the effective application of vision transformer in the prediction of postoperative cancer metastasis still has great research value.

**Acknowledgements** This work is supported by the National Natural Science Foundation of China under Grant No. 61872419, No.61873324, the Natural Science Foundation of Shandong Province, China, under Grant No. ZR2020QF107, No. ZR2020MF137, No. ZR2019MF040, ZR2019MH106, No. ZR2018BF023, the China Postdoctoral Sci-ence Foundation under Grants No. 2017M612178. University Innovation Team Pro-ject of Jinan (2019GXRC015), Key Science & Technology Innovation Project of Shandong Province (2019JZZY010324,2019JZZY010448), and the Higher Educa-tional Science and Technology Program of Jinan City under Grant with No. 2020GXRC057.

## References

1. Rongshou, Z., Chunfeng, Q., Siwei, Z., Hongmei, Z., Kexin, S., Xiuying, G., Changfa, X., Zhixun, Y., He, L., Wenqiang, W., Wangqing, C., Jie, H.: Liver cancer incidence and mortality in China: temporal trends and projections to 2030. *Chin. J. Cancer Res.* **30**(6), 571–579 (2018)
2. Kojiro, M., Nakashima, T.: Pathology of hepatocellular carcinoma. *Nihon Geka Gakkai zasshi* **84**(9), 939–942 (1983)
3. Kim, E., Viatour, P.: Hepatocellular carcinoma: old friends and new tricks. *Exp Mol Med* **52**, 1898–1907 (2020)
4. Hamamoto, I., Okada, S., Hashimoto, T., Wakabayashi, H., Maeba, T., Maeta, H.: Prediction of the early prognosis of the hepatectomized patient with hepatocellular carcinoma with a neural network. *Comput Biol Med* **25**(1), 49–59 (1995)
5. Tsilimigras, D.I., Mehta, R., Moris, D., et al.: Utilizing machine learning for pre-and post-operative assessment of patients undergoing resection for BCLC-0, A and B hepatocellular carcinoma: implications for resection beyond the BCLC guidelines. *Ann Surg Oncol* **27**(3), 866–874 (2020)
6. Chong-Chi, C., King-Teh, L., Hao-Hsien, L., et al.: Comparison of models for predicting quality of life after surgical resection of hepatocellular carcinoma: a prospective study. *J Gastrointest Surg* **22**(10), 1724–1731 (2018)
7. Wen-Hsien, H., King-Teh, L., Hong-Yaw, C., Te-Wei, H., Herng-Chia, C.: Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network. *PLoS One* **7**(1):e29179 (2012)
8. Guoliang, Q., Jun, L., Aiming, H., Zhenlin, Y., Wan-Yee, L., Feng, S.: Artificial neural networking model for the prediction of post-hepatectomy survival of patients with early hepatocellular carcinoma. *J Gastroenterol. Hepatol.* **29**(12), 2014–2020 (2014)
9. Dan, L., Fei, L., Xiaoyan, Xie., Liya, S., et al.: Accurate prediction of responses to transarterial chemoembolization for patients with hepatocellular carcinoma by using artificial intelligence in contrast-enhanced ultrasound. *Eur. Radiol.* **30**(4), 2365–2376 (2020)
10. Jie, P., Kang, S., Zhengyuan, N., et al.: Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging. *Eur. Radiol.* **30**(1), 413–424 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

13. Ze, L., Yutong, L., Yue, C., et al.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ArXiv, abs/2103.14030 (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929 (2020)
15. Jie, H., Li, S., Gang, S.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 2011–2023 (2020)
16. Sanghyun, W., Jongchan, P., Joon-Young, L.: CBAM: convolutional block attention module. In: ECCV (2018)

# Attention-Aware U-Net Network for Segmentation of Retinopathy Region



Wenyang Kong , Fan Li , Ruiwen Xing , Xizhan Gao , Hui Zhao , Jie Su , and Sijie Niu

**Abstract** The important function of the computer-aided analysis platform for OCT (Optical Coherence Tomography) retinal images is image segmentation and the accuracy of the segmentation results affects the diagnosis of the disease area. In this paper, we improve the U-Net network by introducing attention mechanism and residual learning for segmenting retinal image lesions. Based on the original U-Net architecture, we embed the Squeeze-and-excitation module into the residual learning branch of the residual structure to replace the backbone network of U-Net. As a result, our improved method not only retains the advantages of fast convergence of network and Solve the problem of the weakening of the neural network's ability to extract features in the deep level, but also considers the correlation between attention learning channels, enhances useful features, and suppresses noise features. The data set used for training and testing of the method in this paper contains 2944 images from 23 patients diagnosed with ocular CSC (Central Serous Chorioretinopathy). We performed 4 sets of cross-validation on this dataset. According to experimental data, the method proposed in this paper can get better segmentation results. Compared with seven methods, our improved method can achieve competitive segmentation results, which is helpful for clinical disease diagnosis.

**Keywords** Medical image segmentation · U-Net · OCT images · Residual learning · Attention mechanism

## 1 Introduction

Medical images mainly include OCT (Optical Coherence Tomography) [1], CT (Computed Tomography), SPECT (Single Photon Emission Computed Tomography), MRI (Magnetic Resonance Imaging), Ultrasound, and other medical imaging.

---

W. Kong · F. Li · R. Xing · X. Gao · H. Zhao · J. Su · S. Niu (✉)

School of Information Science and Engineering, University of Jinan, Jinan 250022, China  
e-mail: [ise\\_niusj@ujn.edu.cn](mailto:ise_niusj@ujn.edu.cn)

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

At present, medical images are a common diagnosis and treatment tool in clinical practice, and the task of processing medical images [2] is becoming increasingly important. OCT imaging technology is a three-dimensional imaging model developed based on OLCR (Optical Low Coherence Reflectance) [3], which is used to explore the lateral scanning of light beams relative to biological tissues. Because the accuracy of OCT imaging technology can reach the micron level, it has become the standard for clinical diagnosis of retinal diseases year by year. It also has the advantages of real-time imaging, non-destructive tissue, non-invasive, non-contact, and low cost.

In medical image analysis [4], the main form is to look for the lesion area by examining several two-dimensional image slices, which often relies on the professional knowledge of doctors to make judgments. Due to the complex structure of the retina, especially when they are diseased, there will be large topological changes, which requires professional doctors to sketch. However, manual drawing and manual marking are very time-consuming, laborious manual segmentation is easily affected by subjective factors, and the segmentation result cannot be determined. At the same time, the incidence of family eye diseases is increasing year by year, but the number of doctors is limited. Computer technology is used to analyze medical images to realize the auxiliary diagnosis of retinal diseases. The segmentation and three-dimensional reconstruction of organs, tissues and pathological areas can effectively help doctors to conduct quantitative analysis on the pathological areas and related parts, thus effectively improving the accuracy and robustness of the medical-assisted analysis system [5].

The algorithm for segmentation of medical images with OCT has been greatly developed. Most methods cannot directly segment the diseased area and need to obtain retinal layer information. Segmentation result obtained by the existing algorithm is quite different from the image annotated by the expert. Existing algorithms can be divided into three categories: unsupervised segmentation, semi-supervised segmentation, and fully supervised segmentation. Unsupervised approaches include the threshold-based algorithm [6], the level set based approach [7], the Enface Fundus Driven approach [8], and so on. Montuoro et al. [9] proposed an algorithm while the retinal layers could be segmented, the correlation between retinal layers was used to modify the results. Since retinal images often contain a lot of noise, a weakly supervised segmentation method based on prior information is proposed. Wang et al. [10] segmented retinal images by constraining label propagation, but the algorithm relied heavily on key slices Three-dimensional model for segmentation in a specific area was proposed [11]. Fully supervised segmentation methods consist of random forest [12], K-Nearest Neighbor [13], deep learning [15] which can identify pathological areas from medical images. Recently, because deep learning has the advantage of active feature extraction, this is a common method to segment and classify images. Roy [16] proposed the RelayNet model to segment the retinal layer and the lesion area. Fang [17] segmented the retina layer by fusing pattern search and convolutional neural network.

Currently, most existing methods are limited in their ability to capture and characterize lesions based on learning methods. Therefore, to overcome the above problems,

this paper proposes a U-Net neural network segmentation algorithm for retinal image lesion regions by introducing the attention mechanism. Specifically, the attention module is added to the residual network model based on the U-Net neural network [18]. As a result, this model not only retains the advantage of fast convergence speed of the ResNet network and solves the problem of weak learning performance of the deep network, but also obtains the attention between channels through acquiring the relevance of each channel and increase the weight of useful features.

## 2 Related Work

### 2.1 U-Net

Due to the powerful ability of U-Net in medical image segmentation, it has gradually become a well-known algorithm. It proposes the method of data enhancement to further utilize manual annotation information. Ensure that the object category is recognized while considering accurate segmentation and positioning. U-Net network adopts symmetric structure, including down sampling and up sampling. In the subsampling path, network uses the traditional CNN (Convolutional Neural Network) structure; In the up-sampling path, each step involves the expansion of the feature map, and since each convolution operation leads to the loss of boundary pixels, the feature map needs to be clipped. At the last layer, use  $1 \times 1$  convolution to reconstruct the image to the size before down-sampling.

### 2.2 ResNet

ResNet network structure was proposed by He [19] in 2015. This article puts forward the idea of residual learning for the first time. The propagation process of data in the neural network, information is often lost. At the same time, there will be gradient disappearance and gradient explosion. Therefore, deep network cannot get more effective training. The information is not transmitted directly in the ResNet network, but is transmitted to the output through a detour, so that the information is well preserved in the transmission. ResNet network includes several residual modules, among which the residual module is defined as:

$$x_{i+1} = h(x_i) + F(x_i, W_i) \quad (1)$$

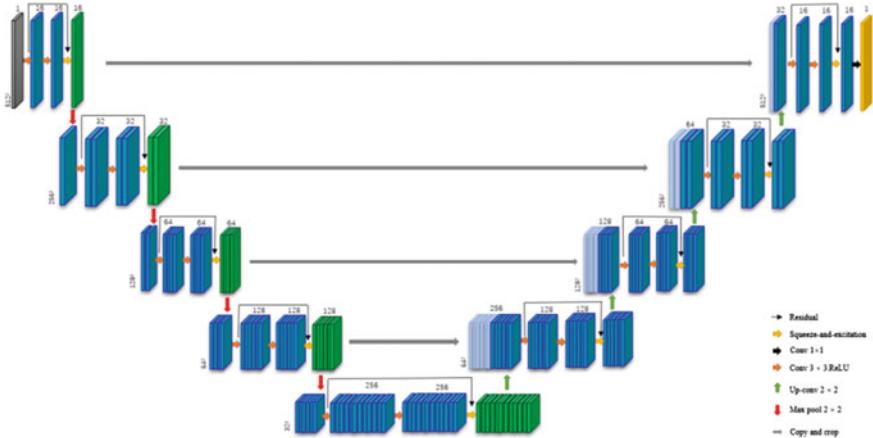
Each residuals module includes direct mapping and residuals.  $x_i$  on the right of Eq. (1) is a direct mapping, and  $F(x_i + W_i)$  represents the residual part. Most of the residuals consist of two or three convolution operations. The ResNet network makes learning much easier by learning the difference between input and output.

## 2.3 SENet

SENet network structure was proposed by Momenta [20] and won first place in ImageNet dataset. SENet builds the model by specifying the relationship between the characteristic channels. In the process of network learning, the importance of each channel is obtained, and then the weight of useful channels is enhanced, and the useless channels of the current task are suppressed to realize the adaptive channel calibration function. The Squeeze operation compresses the feature channel along the spatial dimension and turns it into a number, thus obtaining the global receptive field and ensuring the same number of feature channels between output and input. The operation is like the central gate for RNN. The weight for each channel is expressed using parameters and is used to explicitly model characteristic channel correlation. The Reweight operation uses the weight obtained by Excitation to indicate the importance of the existence of the feature channel, and finally obtains the new feature through weighted multiplication, which realizes the importance of the original feature.

## 3 Methodology

Although encoder-decoder structure is classical in neural networks, it still has great limitations. U-Net can extract important features through convolutional pooling, the receptive field obtained by U-Net convolution is very narrow, and the usual way to increase the receptive field is to deepen the number of network layers. The partition structure of this paper is selected based on the U-Net backbone architecture. In the encoding stage, the maximum pool is used under four sampling, and the samples are decoded by trilinear interpolation High-level semantic encoder is restored to the original image resolution. The high-level features and low-level features in image fusion are fused using jump links to help the decoder recover image details using upper sampling. At the same time, inspired by residual learning and attention mechanism, we embedded the SE module into the residual learning branch of the residual structure to replace the backbone network of the original U-Net. According to the experiment, the model can keep the advantages of fast convergence of ResNet, and consider the correlation between SENet learning channels, the attention to the channel is shielded, the useful features will be focused on, and useless features will be ignored. The input image of this network structure is  $128 \times 512 \times 512$ . In addition to  $2 \times 2 \times 2$  convolution kernel size,  $2 \times 2 \times 2$  step maximum pooling operation, all convolutions use  $3 \times 3 \times 3$  convolution kernel,  $1 \times 1 \times 1$  step size and  $1 \times 1 \times 1$  filling. Meanwhile, batch normalization layer and ReLU activation function are added after each convolution layer. The loss function is the cross-entropy function, the formula is as follows:



**Fig. 1** SERes-UNet network structure

$$L = -\frac{1}{N} \sum_i y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (2)$$

where  $y_i$  represents the label of the original data,  $p_i$  represents the probability of the predicted positive sample.

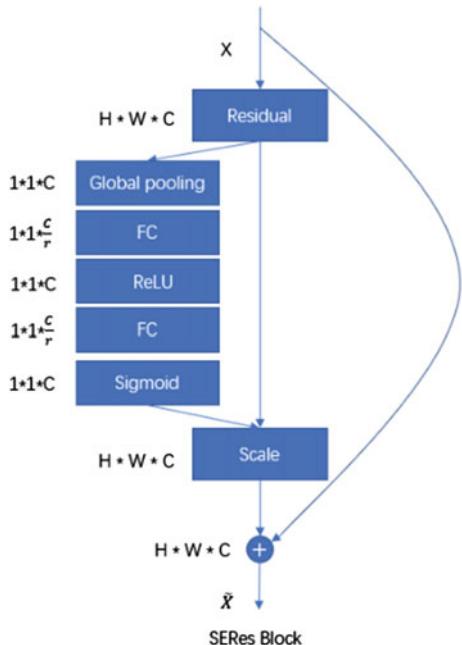
Specifically, during encoding and decoding, the input of each layer goes through two convolution operations, and the feature maps after the two convolutions are globally averaged and pooled. Then two full-connection layers are connected and Add the sigmoid function to activate after the output of the second fully connected layer, and the value is multiplicatively to the feature maps output by the two convolutions. It is added to the input of this layer pixel by pixel. In each layer, jump connection is used to join the channel dimensions of the feature graphs of the same size when encoding and decoding. The network structure is shown in Fig. 1.

The SERes block is shown in Fig. 2. First, we do a global average pooling for the multiple feature graphs output from the remainder structure. The Squeeze process is the  $1 \times 1 \times C$  data output. At last, use sigmoid to limit the output to the range of [0, 1], Then multiply the obtained values with the  $C$  channel of the residual structure, and output them as the input data of the next stage.

## 4 Experimental Results

### 4.1 Dataset

The data set used in this study was composed of 23 patients with retinopathy. Each patient included 128 images with an image size of  $512 \times 1024$ , and a total of 2944

**Fig. 2** SERes block

$512 \times 512$  images were obtained after background removal. In the actual training and test, the 23 patients were divided into 4 groups by 4 times of cross validation. Use three groups as the training set and the latter group as the test set.

## 4.2 Evaluation Metrics

In this paper, Dice coefficient, Positive Predictive Value (PPV) and Sensitivity were selected as indicators to measure segmentation results [21]. Dice coefficient was first named after Lee Raymond Dice, and later it was widely used as a measure of image segmentation algorithms. Dice coefficient can measure the similarity of two sets. Its maximum value is 1 and the minimum value is 0. The larger the value is, the better the segmentation result is, as defined below:

$$Dice(A, B) = \frac{2(A \cap B)}{(A + B)} \quad (3)$$

A and B represent two sets respectively, the numerator represents the intersecting part of the two sets, and the denominator represents the sum of the two sets. Dice reflects the good or bad segmentation results by comparing the amount of overlap between two sets in the sum of the two sets.

PPV refers to the proportion of correct prediction in the data with positive prediction results, and is defined as follows:

$$PPV = \frac{TP}{(TP + FP)} \quad (4)$$

where TP is the value that is judged correctly in the predicted positive sample, and FP is the value that is judged incorrectly in the predicted positive sample.

Sensitivity represents the proportion of predicted positive samples to actual positive samples, as defined below

$$SEN = \frac{TP}{(TP + FN)} \quad (5)$$

FN means it is predicted to be negative, but it is positive sample.

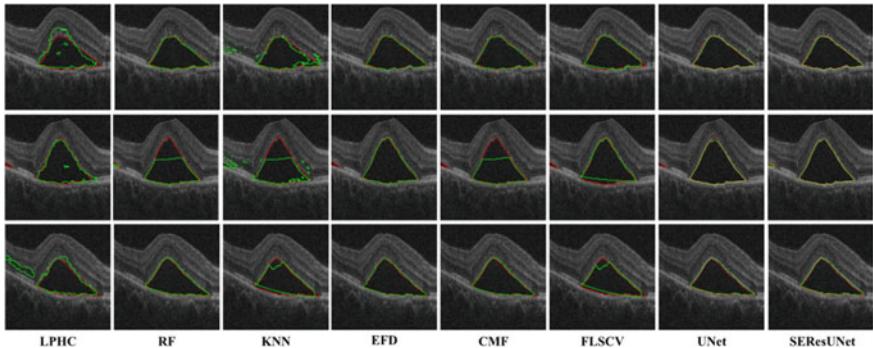
### 4.3 Experimental Results

To verify the performance of the proposed method, LPHC [10], FR [12], KNN [13], EFD [8], CMF [11], FLSCV [7] and UNET [18] are compared. By comparing a variety of segmentation algorithms, the method we proposed get a good segmentation effect on retinal OCT images, and has achieved good segmentation results under the three evaluation criteria. As shown in Table 1, quantitative comparison of various methods shows the method we proposed can achieve more accurate than other methods of segmentation, including SEN index reached 95.8%, variance control at 2%. Compared with the traditional U-Net method, after introducing the attention mechanism and residuals, the proposed method can better capture the characteristics of the lesion area, thus improving the segmentation results.

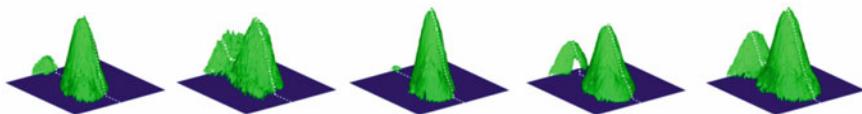
Figure 3 visualizes the segmentation results of various segmentation algorithms, where the red solid line represents the real label used for network training, and the green solid line represents the segmentation results of different methods. We can get

**Table 1** The results of different method

	SEN	Dice	PPV
LPHC [10]	$81.3 \pm 9.4$	$65.6 \pm 10.4$	$55.6 \pm 13.3$
RF [12]	$92.6 \pm 4.4$	$87.1 \pm 4.3$	$92.4 \pm 2.0$
KNN [13]	$80.9 \pm 6.6$	$86.1 \pm 4.1$	$91.9 \pm 3.8$
EFD [8]	$94.2 \pm 5.2$	$93.0 \pm 4.8$	$93.7 \pm 4.0$
CMF [11]	$92.1 \pm 4.1$	$93.0 \pm 3.4$	$93.9 \pm 2.5$
FLSC V[7]	$84.4 \pm 20.4$	$86.2 \pm 7.3$	$78.9 \pm 21.7$
UNet [18]	$92.8 \pm 2.4$	$91.6 \pm 3.4$	$90.6 \pm 3.9$
SERes-UNet (ours)	<b><math>95.8 \pm 2.0</math></b>	<b><math>93.1 \pm 3.0</math></b>	$91.2 \pm 5.6$



**Fig. 3** Segmentation results of different algorithms



**Fig. 4** D rendering of segmentation results of our method

the conclusion that the segmentation by the method in this paper is close to the actual label, and the contour is clear and smooth, almost coincides with the label. However, most methods cannot guarantee the smoothness of the contour, and there are often different degrees of under segmentation or over segmentation. At the same time, we performed 3D visualization [22] of the segmentation result as shown in Fig. 4 to facilitate further observation of the segmentation effect.

## 5 Conclusion

In this paper, we embed the attention module to the residual learning branch to replace the original U-Net backbone and perform automatic segmentation of retinal OCT images. A data set composed of 23 patients with retinopathy was verified. We can know the proposed method can retains the advantage of fast network convergence and gets the attention between channels by obtaining the relevance of each channel, enhance the weight of useful features and suppress the weight of useless features. After cross-validation of data set images, the sensitivity, DICE coefficient and positive predictive value reached 95.8, 93.1 and 91.2%, which achieved better segmentation results compared with other segmentation algorithms.

**Acknowledgements** This work is supported by the National Natural Science Foundation of China under Grant No. 61872419, No. 61873324, the Natural Science Foundation of Shandong Province, China, under Grant No. ZR2020QF107, No. ZR2020MF137, No. ZR2019MF040,

ZR2019MH106, No. ZR2018BF023, the China Postdoctoral Science Foundation under Grants No. 2017M612178. University Innovation Team Project of Jinan (2019GXRC015), Key Science & Technology Innovation Project of Shandong Province (2019JZZY010324,2019JZZY010448), and the Higher Educational Science and Technology Program of Jinan City under Grant with No. 2020GXRC057.

## References

1. Yankui, S.: Medical image processing and application of optical coherence tomography. *Opt. Precision Eng.* **22**(04), 1086–1104 (2014)
2. Qian, Z.: Research and Application of Medical Image Segmentation. Southern Medical University (2014)
3. Boehnke, M., Masters, B.R., Waelti, R., et al.: Precision and reproducibility of measurements of human corneal thickness with rapid optical low-coherence reflectometry (OLCR). *J. Biomed. Opt.* **4**(1), 152–157 (1999)
4. Yizhou, Y., Dejun, S., Jiechao,M., etc al.: Application progress of artificial intelligence in medical image analysis. *Chin. Med. Imag. Technol.* **35**(12), 1808–1812 (2019)
5. Ruohan, Z., Yang, G., Kui, S., et al.: Design and application of medical image computer aided analysis system. *Software* **40**(10), 68–72 (2019)
6. Wilkins, G.R., Houghton, O.M., Oldenburg, A.L.: Automated segmentation of intraretinal cystoid fluid in optical coherence tomography. *IEEE Trans. Biomed. Eng.* **59**(4), 1109–1114 (2012)
7. Wang, J., Zhang, M., Pechauer, A.D., et al.: Automated volumetric segmentation of retinal fluid on optical coherence tomography. *Biomed. Opt. Express* **7**(4), 1577–1589 (2016)
8. Menglin, W., Qiang, C., Xiaojun, H., et al.: Automatic subretinal fluid segmentation of retinal SD-OCT images with neurosensory retinal detachment guided by enface fundus imaging. *IEEE Trans. Biomed. Eng.* **65**(1), 87–95 (2017)
9. Alessio, M., Sebastian, M.W., Bianca, S.G., et al.: Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context. *Biomed. Opt. Express* **8**(3), 1874–1888 (2017)
10. Tao, W., Zexuan, J., Quansen, S., et al.: Label propagation and higher-order constraint-based segmentation of fluid-associated regions in retinal SD-OCT images. *Inf. Sci.* **358**, 92–111 (2016)
11. Menglin, W., Wen, F., Qiang, C., et al.: Three-dimensional continuous max flow optimization-based serous retinal detachment segmentation in SD-OCT for central serous chorioretinopathy. *Biomed. Opt. Express* **8**(9), 4257–4274 (2017)
12. Andrew, L., Aaron, C., Emily, K.S., et al.: Automatic segmentation of microcystic macular edema in OCT. *Biomed. Opt. Express* **6**(1), 155–169 (2015)
13. Quellec, G., Lee, K., Dolejsi, M., et al.: Three-dimensional analysis of retinal layer texture: identification of fluid-filled regions in SD-OCT of the macula. *IEEE Trans. Med. Imaging* **29**(6), 1321–1330 (2010)
14. Shi, X., Luo, S.: Support vector machine based medical image segmentation. In: Proceedings of the 2010 International Conference on Image Processing, Computer Vision & Pattern Recognition, pp. 402–405. Las Vegas NV (2010)
15. Venhuizen, F.G., van Ginneken, B., Liefers, B., et al.: Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. *Biomed. Opt. Express* **9**(4), 1545–1569 (2018)
16. Roy, A.G., Conjeti, S., Karri, S.P.K., et al.: ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomed. Opt. Express* **8**(8), 3627–3642 (2017)

17. Leyuan, F., David, C., Chong, W., et al.: Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed. Opt. Express* **8**(5), 2732–2744 (2017)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham (2015)
19. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society (2016).
20. Jie, H., Shen, L., Samuel, A., et al.: Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 2011–2023 (2017)
21. Chunyan, L.: Research on Evaluation Method of Image Segmentation. Xi'an University of Electronic Science and Technology (2011)
22. Luping, F., Guopeng, L., Wenjie, H., et al.: Research on 3D visualization of medical image based on WebGL. *Computer Syst. Appl.* **22**(9), 25–30 (2013)

# LESN: Low-Light Image Enhancement via Siamese Network



Xixi Nie , Zilong Song , Bing Zhou , and Yating Wei

**Abstract** Images have become an important medium for people to get in touch with the world and new things. Low-light images enhancement is often used in target detection, night reconnaissance, government office and other fields. Images collected in low-light environment suffer from insufficient brightness, which makes it difficult for the images to be effectively used in subsequent vision tasks. In response to this problem, many researchers have conducted in-depth research on low-light image enhancement. However, their methods mostly failed in extremely low-light scenes, and even amplified the underlying noise in the images. That is why we propose a new low-light image enhancement method that integrates deep reinforcement learning via siamese network. We conducted experiments on the public LOL dataset and achieved good experimental results.

**Keywords** Deep reinforcement learning · Siamese network · Image enhancement · Low-light image

## 1 Introduction

From security applications, government offices and commercial recommendation systems, high-quality images are the key to automated and humanized decision-making. The high-quality images are captured by the camera system, providing sufficient evidence for the action process. Visual information in a dynamic environment is collected and accurately processed the data that is critical to making informed decisions and ensuring the mission success. However, images taken in a dark environment will have many problems such as poor visibility, low contrast, high noise, loss of detail, and color distortion. Although automatic exposure mechanisms (such as ISO) can enhance image brightness, they also have passive effects (such as blur, over saturation).

---

X. Nie · Z. Song · B. Zhou · Y. Wei  
Zhengzhou University, Zhengzhou Henan 450000, China  
e-mail: [xixinie2019@gs.zzu.edu.cn](mailto:xixinie2019@gs.zzu.edu.cn)

In the past, people have proposed many methods to solve the above problems. In summary, it mainly including three categories: Spatial methods are directly processing pixels, such as histogram equalization. Pizer [1] proposed adaptive Histogram Equalization Algorithm. But this method increases the image noise while increasing the brightness. The frequency domain method operates in a certain transform domain, such as wavelet transform. Hybrid domain method is a combination of some methods in the space domain and frequency domain. But, they only focus on improving the contrast and brightness of the images, while ignoring the impact of noise and the contextual information in the images, and even causing noise amplification. Recently, image enhancement is a prominent application in deep learning. Wang [2] proposed a method of using neural network to estimate and adjust the illumination layer of low-light images. Ma [3] converts the color space and performs the brightness channel Enhancement, so as to obtain a normal exposure image. Zhang [4] proposed the HSV color space conversion RetinexNet low-light image enhancement method. Although significant progress has been made, most of these methods perform poorly in an extremely low-light environment, with problems such as noise, loss of details and overexposure.

Therefore, we propose a new low-light image enhancement method that integrates deep reinforcement learning via siamese network. The method introduces the siamese network and reinforcement learning into itself. Through multiple rounds of iterations, an excellent method is obtained. The user-satisfied enhanced images are output. The method was proved effectively solving above problems. At the same time, the network converges faster, having strong generalization ability. SSIM is better than other comparison methods. The obtained low-light enhancement images are more natural. Our main contributions are as follows:

- (1) We introduce Siamese network to the field of low-light image enhancement, and propose a new deep reinforcement learning low-light image enhancement method that integrates the Siamese network.
- (2) We integrate deep reinforcement learning, and propose an end-to-end low-light image enhancement method.
- (3) Our network has a astrong generalization ability. And the objective evaluation index is better than other comparison methods.

## 2 Related Work

### 2.1 Low-Light Image Enhancement

Low-light image enhancement is used to enhance the visual quality of low-light images, improve the visibility of image details, and increase the signal-to-noise ratio. In recent years, many researchers have been working on low-light images. Among them, the better effect is the histogram equalization algorithm [5]. The histogram of the output image is suppressed to satisfy the relevant constraints. Pizer [1] proposed

an adaptive histogram equalization algorithm, but this method increases image noise while increasing brightness.

In addition, there are also many algorithms that apply retinex theory [6] to this field. This theory assumes, for colorful images, which can be decomposed into two parts in reflectance and illuminance. The earliest method to apply this theory is the Single Scale Retina (SSR) [7], which limits the smoothing of the illumination map through a Gaussian filter. Multiscale Retina (MSRCR) [8] extends SSR with multiscale Gaussian filters and color restoration. LIME [9] uses only structural prior knowledge for illuminance estimation and uses reflections as the final result.

Neural networks have become a prominent mean of image enhancement with the continuous development of machine learning. Gharbi [10] proposed a bilateral learning framework for enhancing low-light images. The Lighten-Net network proposed by Li [11] combines Retinex theory with convolutional layers to enhance low-illuminance images. Zhang [4] proposed a Retinex-Net low-light image enhancement algorithm for HSV color space conversion. Xu [12] proposed a low-light residual convolutional network. The network can denoise while enhancing the contrast. Although they have made these work, these methods perform poorly in low-light environment.

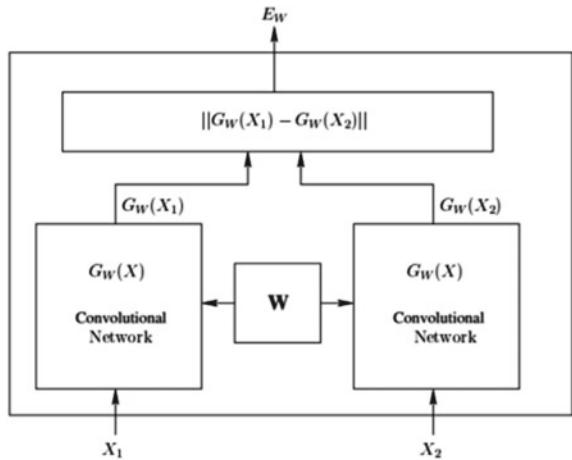
## 2.2 *Siamese Neural Network*

Siamese neural network is based on a coupled architecture established by two artificial neural networks. The outstanding feature of the siamese network is the sharing of weights. It is a type of supervised learning, used for metric learning. Under the supervised learning paradigm, siamese network maximizes the representation of different labels and minimizes the representation of the same label. Prior to 2015, the Siamese network was popular for image processing. Since then, the siamese network has evolved further tracking fields. Figure 1 shows a typical model of Siamese network.

## 2.3 *Reinforcement Learning*

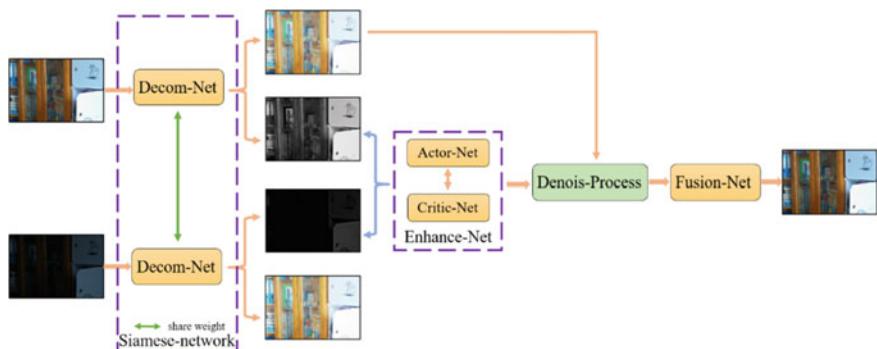
Reinforcement learning is an important branch of machine learning. It, in recent years, has been widely used in image enhancement [13–15]. By receiving rewards for actions, it can directly obtain learning information and update model parameters. Park [13] proposed a deep enhancement learning method for color enhancement. It treats color enhancement as a Markov decision process, and trains the agent to learn the best global enhancement order at each step.

**Fig. 1** The basic schematic diagram of the Siamese network



### 3 Framework

In this paper, we propose a new low-light image enhancement model that incorporates deep reinforcement learning via siamese network (Fig. 2). The model mainly includes four parts: image separation based on siamese network, image enhancement network, denoising processing, and image fusion network. First, we input a set of image pairs into the siamese network, and use Retinex theory to separate the images. The separated image contains content information image and lighting information image. We enhance the low-light image. And the core work is to enhance the light information images. Reinforcement learning method is used to enhance the illumination information images. But the enhanced images are easy to introduce noise, we will denoise the image. Finally, image fusion is performed to restore the original color information of the images to obtain more natural images.



**Fig. 2** The framework of low-light image enhancement based on deep reinforcement learning via siamese network

### 3.1 Retinex Theory and Siamese Network

Retinex uses an image enhancement method based on scientific analysis. The color of the object is not affected by the unevenness of lighting. It is based on color consistency (color constancy). Unlike traditional methods, which can only enhance certain types of image features, retinex can achieve a balance of dynamic range compression, edge enhancement, and color constancy. Therefore, it can be adapted to enhance different types of images.

Taking advantage of the features of retinex theory, we propose a new low-light image enhancement method that combines deep reinforcement learning with siamese network. The two input branches of the Siamese network input low-quality images and high-quality images respectively. Then they would be separated into images containing contextual and optical information via Decom-net. The principle is as follows:

$$S_{(x,y)} = R_{(x,y)} \cdot L_{(x,y)} \quad (1)$$

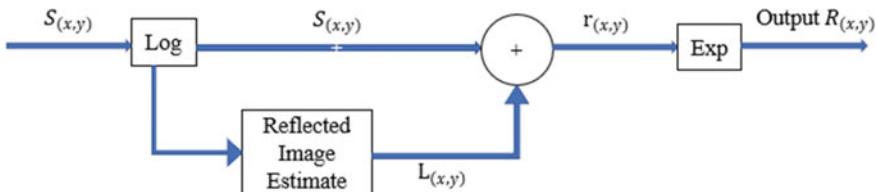
where,  $L_{(x,y)}$  represents the incident light image.  $R_{(x,y)}$  represents the reflection property image of the object.  $S_{(x,y)}$  represents the reflected light image that the human eye can receive.

The processing process of the Retinex algorithm is as follows (Fig. 3).

Normally, we assume that the illuminated image is estimated to be a spatially smooth image, the original image is  $S_{(x,y)}$ , the reflected image is  $R_{(x,y)}$ , and the brightness image is

$$r_{(x,y)} = \log R_{(x,y)} = \log \frac{S_{(x,y)}}{L_{(x,y)}} \quad (2)$$

Decom-net takes low-quality images and high-quality images as input, and estimates the reflectance and illuminance of the high-quality and low-quality images respectively. A  $3 \times 3$  convolutional layer is used to extract features. Then several  $3 \times 3$  convolutional layers are used, and the PReLU unit is used as the activation function to map the RGB image into the reflection map as well as the illumination map. Decom-net separates low-quality images from high-quality images by sharing



**Fig. 3** Retinex algorithm flow chart

weights, and obtains images containing content information and images containing lighting information.

Obviously, for the images shot in the same scene, the content information image is not affected by the illumination factor, so the separated images containing the content information are completely consistent, and the discrimination lies in the difference in the illumination information. Therefore, our object is to enhance the brightness map. To effectively solve this problem, we introduce reinforcement learning into low-light image enhancement, and complete the image enhancement work efficiently and quickly.

### 3.2 Image Enhancement

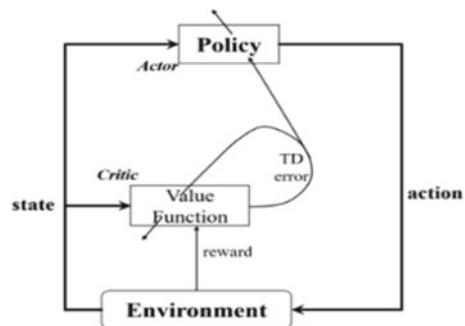
In order to automatically obtain the enhanced image in accordance with the natural lighting environment, we propose a low-light image quality enhancement algorithm, which introduces reinforcement learning into the low-light image enhancement process.

When we use retouching software to adjust the illumination information of the image, we will modify the low-light image according to the image light information under natural light conditions, which can be regarded as a problem of sequential decision-making. Inspired by this, we use the Actor-Critic strategy to train the low-light image enhancement model. The Actor-Critic strategy is shown in Fig. 4.

In the above figure, the agent chooses action according to the strategy. This critic network updates the value function according to the rewards given by the environment, and promotes the actor network to choose a better strategy. As the number of iterations increases, the actor obtains a reasonable probability for each action, and the critic continues to increase the reward value of the action in each state.

We describe this process as  $P = (S, A)$  to enhance the low-light images. In the reinforcement learning framework,  $S$  represents the state space, and  $A$  represents the action space, which is a collection of a series of filtering operations.

**Fig. 4** Reinforcement learning block diagram



As we learned [16], when the images are enhanced, the filter parameters need to be determined. The action space consists of two parts: the discrete set of filter  $a_1$  and the continuous set of filter parameters  $a_2$ .  $a_1$  contains 8 filtering operations that reflect the characteristics of image lighting, and  $a_2$  has different continuous values according to the differences of each filter.

$$a_1 = \left\{ \begin{array}{l} \text{gamma, exposure, contrast, tone, color} \\ \text{curve, whitebalance, saturation, WNB} \end{array} \right\} \quad (3)$$

Therefore, the actor selection strategy contains two parts:  $\pi = (\pi_1, \pi_2)$ ,  $\pi_1$  represents the probability distribution of each filter, and  $\pi_2$  is the filter parameter generated by the actor network. The critic network evaluates the generated results.

Our actor-critic network consists of 5 convolutional layers and 1 fully connected layer, and the output fully connected layer is 128 dimensions. The network structure is shown in the Fig. 5.

The expectations of the critic network are:

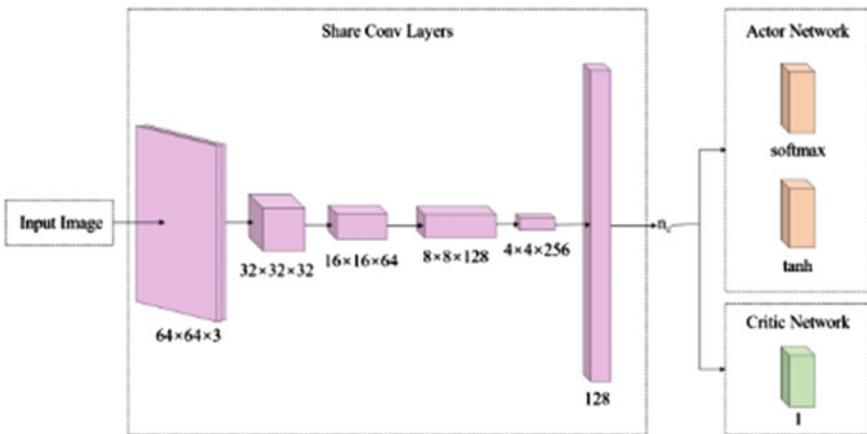
$$V_\pi(s) = E_\pi[r + \gamma V_\pi(s')] \quad (4)$$

To evaluate this strategy, our value function is defined as follows:

$$Q_\pi(s, a) = R_s^a + \gamma V_\pi(s') \quad (5)$$

The loss function of the critic network is defined as follows:

$$L_c = \frac{1}{n} \sum_{i=1}^n [A_\pi(s, a)]^2 \quad (6)$$



**Fig. 5** Actor/critic network architecture

where,  $A_\pi(s, a)$  is the Monte Carlo estimation of the optimization function.

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s) = r + \gamma V_\pi(s') - V_\pi(s) \quad (7)$$

The Actor network selects actions based on the strategy  $\pi$  to obtain more return values. The goal of the Actor network is to maximize  $L_a$ .

$$L_a = -\frac{1}{n} \sum_{i=1}^n A_\pi(s, a) \log \pi(s, a) \quad (8)$$

### 3.3 Image Denoising and Fusion

After image separation and enhancement, the image will be affected by a variety of noises, such as Gaussian noise, salt and pepper noise. Dabov [17] proposed the BM3D denoising algorithm. That is an image noise reduction method, which has improved the sparse representation of the image in the transform domain. The advantage of the BM3D noise reduction method is to better retain some details in the image. BM3D uses different noise reduction strategies. By searching for similar blocks and filtering in the transform domain, the block evaluation value is obtained, and finally each point in the image is weighted to obtain the final denoising effect.

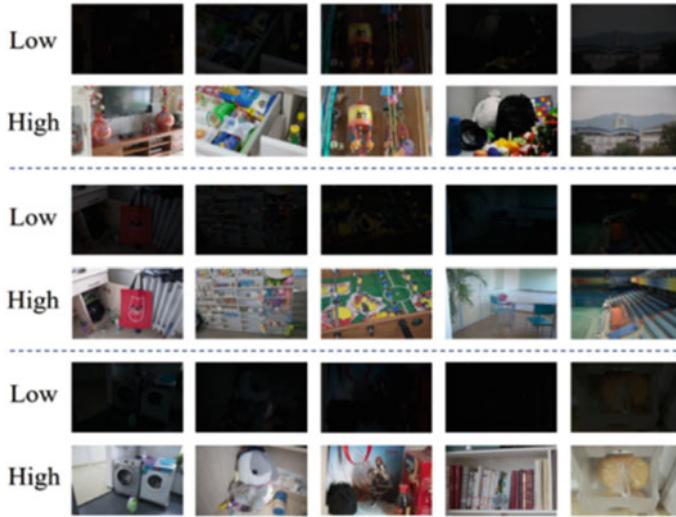
The image after denoising is still black and white, lacking the characteristics of the image itself. Therefore, we need to color the enhanced and denoised image according to the image content information, and output a color image that basically conforms to natural lighting.

## 4 Experiment

### 4.1 Dataset

Although this field has a long history of research, there are few paired low-light level image data sets captured in real scenes. Some work uses high dynamic range datasets as alternatives. However, these datasets include a few images about scale and contain limited scenarios. Therefore, in order to facilitate the learning of low-light enhancement networks from large-scale datasets, we use real photographic dataset and dataset from raw image synthesis. The former is mainly used to capture image features and attributes in real situations. And the latter is mainly used to diversify scenes and objects.

LOL (Low Light Pair Dataset) is the first image dataset used to enhance low-light images in a real scene. It contains 500 low/normal light image pairs. Most low



**Fig. 6** Some images in natural scenes

light images are collected by changing ISO, but other camera configurations are fixed. Images from this dataset are taken from a variety of scenes, including homes, campuses, clubs, and streets. Figure 6 shows a subset of the scene.

In addition, we used a synthetic dataset. We analyzed the illuminance distribution of the low-light image. We collected 270 low-light images from public MEF [18], NPE [19], LIME [9], DICM [20] and Fusion [21] datasets, converted the images into YCbCr channels, and calculated the histogram of the Y channel. We also collected 1000 original images from RAISE [22] as natural light images.

## 4.2 Experimental Details

The LOL dataset containing 500 image pairs mentioned in the third part is decomposed into two parts, of which 485 pairs of images are used for training, and the other 15 pairs of images are used for evaluation. Therefore, the network is trained on the basis of 485 pairs of real images and 1000 pairs of synthetic image sets. The entire network is light-weight. Decom-net uses 5 convolutional layers. And there is PReLU activation between two convolutional layers without PReLU.

The enhancement network is composed of two actor-critic networks related to reinforcement learning. Better results can be obtained in the case of small sample learning. We first train the decomposition network and the enhancement network. Then we use the gradient descent method to fine-tune the network. The batch size is 16. Patch size is  $48 \times 48$ .

**Table 1** Our experimental results based on the LOL dataset

Methods	PSNR	SSIM
HE	15.81	0.5607
MSR	16.69	0.5262
NPE	<b>16.97</b>	0.5894
LIME	16.76	0.5644
Retinex-Net	16.77	0.5594
Ours	15.49	<b>0.6729</b>

The bold values represent the PSNR and SSIM scores in different models

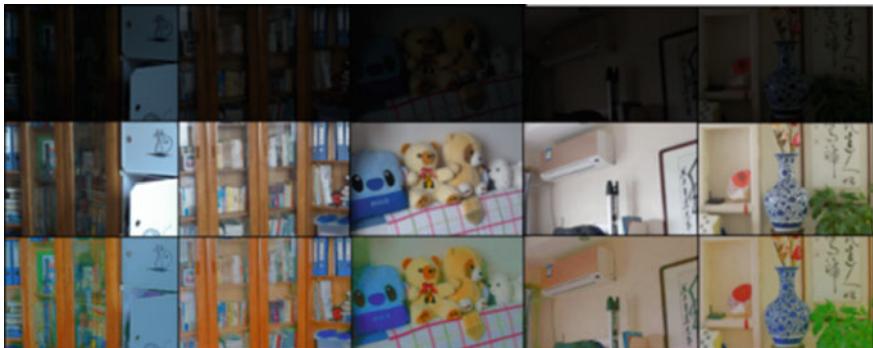
### 4.3 Experimental Environment

We use Windows10 Professional operation system. The configuration parameters are as follows: NVIDIA-RTX3000 professional 3D graphics card, 10th generation Intel Core i9 eight-core processor. CUDA version is 11.0. Tensorflow-GPU version is 2.4.1.

### 4.4 Experiment Result

We evaluated our model in images of real scenes, and compared it with related models on the same dataset. The experimental results are shown in Table 1.

It can be seen that our experiment has achieved good results. The images of some experimental results are shown in Fig. 7.



**Fig. 7** The first line shows the low-light images. The second line shows the images under natural lighting. The third line shows the results of our experiment

## 5 Conclusion

We propose a new low-light image enhancement method that integrates deep reinforcement learning via a sham network. By this method, the observed image can be decomposed into reflectance and illuminance in a data-driven manner without decomposing the reflectance and illuminance. Then we use reinforcement learning to enhance the contrast image and denoise the reflectance. Decomposing the network and enhancing the network are end-to-end training. The experimental results show that our method produces a visually satisfactory effect. From the data point of view, our experimental results have made significant progress. However, it can be seen from our experimental results that there are still noise effects. Therefore, we will focus on solving this problem in the future.

## References

1. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *Pattern Anal. Mach. Intell.* **33**(12), 2341–2353 (2011)
2. Wang, W., Wei, C., Yang, W.: GLADNet: low-light enhancement network with global awareness. In: *IEEE International Conference on Automatic Face Gesture Recognition*, pp. 751–755 (2018)
3. Ma, H.Q., Ma, S.P., Xu, Y.L.: Low-light image enhancement based on deep convolutional neural network. *Acta Optica Sinica* **39**(2), 0210004 (2019)
4. Zhang, H. Y., Zhao, J. D.: RetinexNet low illumination image enhancement algorithm in HSV Space. *Laser Optoelectron. Progr.* **57**(20), 201504 (2020)
5. Stephen, M., Pizer, E., Philip, A., John, D.A., Karel, Z.: Adaptive histogram equalization and its variations. *Computer Vision Graph. Image Process.* **39**(3), 355–368 (1987)
6. Edwin, H.L.: The retinex theory of color vision. *Sci. Am.* **237**(6), 108 (1977)
7. Daniel, J.J., Zia-ur, R., Glenn, A.W.: Properties and performance of a center/surround retinex. *IEEE Trans. Image Process.* **6**(3), 451–462 (1997)
8. Daniel, J.J., Zia-ur, R., Glenn, A.W.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **6**(7), 965–976 (1997)
9. Xiaojie, G., Yu, L., Haibin, L.: Lime: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **26**(2), 982–993 (2017)
10. Gharbi, M., Chen, J., Barron, J.T.: Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph.* **36**(4) (2017)
11. Li, C., Bimef, F.J., Porikli, F.: LightenNet: a convolutional neural network for weakly illuminated image enhancement. *Pattern Recogn. Lett.* **104**, 15–22 (2018)
12. Xu, W., Lee, M., Zhang, Y.: Deep residual convolutional network for natural image denoising and brightness enhancement. In: *International Conference on Platform Technology and Service*. IEEE (2018).
13. Jongchan, P., Joon-Young, L., Donggeun, Y., In-So, K.: Distort-and-recover: Color enhancement using deep reinforcement learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5928–5936 (2018)
14. Jianzhou, Y., Stephen, L., Sing-Bing K., Xiaoou, T.: A learning-to-rank approach for image color enhancement. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2987–2994 (2014)
15. Yuanming, H., Hao, H., Chenxi, X., Baoyuan, W., Stephen, L.: Exposure: A white- box photo post-processing framework. *ACM Trans. Graph. (TOG)* **37**(2), 26 (2018)

16. Yuanming, H., Hao, H., Chenxi, X., Baoyuan, W., Stephen, L.: Exposure: a white-box photo post-processing framework. *ACM Trans. Graphi. (TOG)* **37**(2), 26 (2018)
17. Hang, Z., Orazio, G., Iuri, F., Jan, K.: Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imag.* **3**(1) (2017)
18. Keda, M., Kai, Z., Zhou, W.: Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.* **24**(11), 3345 (2015)
19. Shuhang, W., Jin, Z., Haimiao, H., Bo, L.: Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Trans. Image Process.* **22**(9), 3538–3548 (2013)
20. Chulwoo, L., Chul, L., Chang-Su, K.: Contrast enhancement based on layered difference representation. In: *IEEE International Conference on Image Processing*, pp. 965–968 (2013)
21. Haomiao, J., Qiyuan, T., Joyce, F., Brian, A.W.: Learning the image processing pipeline. *IEEE Trans. Image Process.* **26**(10), 5032–5042 (2017)
22. Duc-Tien, D., Cecilia, P., Valentina, C., Giulia, B.: Raise: a raw images dataset for digital image forensics. In: *ACM Multimedia Systems Conference*, pp. 219–224 (2015)

# A Lightweight-Improved CNN Based on VGG16 for Identification and Classification of Rice Diseases and Pests



Kaibo Liang , Yuzhi Wang , Li Sun , Dongpeng Xin , and ZiWei Chang

**Abstract** In view of the fact that there are too many parameters in the process of identifying and classifying rice pests and diseases by using convolution neural network, so it is difficult to be applied in mobile terminals. In this paper, based on VGG16, the number of channels is reduced to lessen the model parameters, and then the depth separable convolution is introduced to replace the original convolution operation of the model in order to further reduce the model parameters. In addition, SE module from SENet is introduced to enhance the feature extraction ability of this model, so as to the lightweight-improved VGG16 is built for the first time, which is named VGG-DS. We introduce VGG16, InceptionV3, ResNet50, MobileNet and other mainstream models for comparison. The experimental results show that the parameters of VGG-DS are only 826,753, and the accuracy reaches 93.66%, which has advantages over other networks. After the hyperparameter of VGG-DS is adjusted, Adamax is selected as the optimizer, and the learning rate is 0.01, the accuracy reaches 95.09%, which is optimal accuracy. VGG-DS can be applied in mobile terminals to the identification and classification of rice diseases and insect pests.

**Keywords** Rice diseases and insect pests · VGG16 · Deep separable convolution · SE module

---

K. Liang · Y. Wang · L. Sun   
Beijing Wuzi University School of Information, Beijing, China  
e-mail: [slsally@163.com](mailto:slsally@163.com)

D. Xin  
Beijing Wuzi University School of Logistics, Beijing, China

Z. Chang  
School of Information, Beijing Wuzi University, Tianjin, China

## 1 Introduction

With the development of the theory of precision agriculture [1], many emerging technologies are continuously integrated with the agricultural field, which greatly promotes the fine operation process of agricultural production and pest control. As one of the main food crops in China, rice occupies a very important position in the field of agricultural production [2]. In 2020, China's total rice output is 21.86 million tons, accounting for 28.9% of the world's rice output, ranking first in the world [3]. Rice diseases and pests will not only lead to the reduction of rice yield, but also cause huge loss of property to farmers. Therefore, timely and effective classification of rice diseases and pests, targeted development of control programs is essential. In recent years, thanks to the substantial improvement of hardware level, computer vision related research has been greatly developed. As an important branch of computer vision, CNN has been widely applied in agriculture field, especially in the field of crop pest identification and classification.

But, the mainstream convolutional neural network models have large parameters and complex model structure, which cannot be applied to terminal equipment with weak computational power. In addition, classification of rice diseases and pests in the scene is often in the field, which leads to the difficulty of effective deployment of large-scale equipment, it is a contradiction between terminal equipment and models. Therefore, it is urgent to reduce the model parameters while ensuring the training effect of the model, so as to better match the mobile terminal equipment and complete classification of rice diseases and pests. Based on this, from the perspective of lightweight network, an improved VGG16 is proposed, which attains a better performance than many state-of-the-art models in the respect of identification and classification of rice pests. The specific contributions of this paper are as follows:

- (1) Lightweight-improved VGG16: A new lightweight CNN model is proposed based on VGG16, which has fewer parameters than several state-of-the-art models by reducing the convolution channels of VGG16 and replacing the original convolution layer with the depth separable convolution and adding SE module in the new model.
- (2) Compare the traditional convolutional neural network and lightweight network: By using the small sample data set to train the traditional convolutional neural network and the lightweight network, the experimental results show that when the data set is small, the traditional convolutional neural network model is difficult to obtain a good result because of many parameters and complex model structure. However, the lightweight network can achieve better results when training small sample data sets by virtue of its own volume advantage.
- (3) Test: The proposed scheme is evaluated through extensive experiments. By comparing with some established CNN models such as MobileNet, Xception, VGG16, InceptionV3, ResNet50 and ResNet101, we evaluate the value of our method in image quality and classification accuracy.

The remanent parts of this paper are structured as follows. Section 2 gives a short overview of the related work. Section 3 introduces the materials and methods,

which contain respectively experimental environment, data sets required for experiments and the overview of the CNN models. In Sect. 4, we concentrate on the new lightweight model. Some extensive simulation results are shown in Sect. 5. Section 6 gives the conclusion and a view of the future works.

## 2 Related Works

Scholars have done a lot of exploration and research on crop diseases and pest recognition based on convolution neural network.

Zhang et al. [4] greatly reduced the model parameters through improved GoogleNet and Cifar10 based on deep learning, and optimized the model by adjusting hyperparameter and pooling type. Agarwal et al. [5] proposed a CNN network with 8 hidden layers because of the excessive number of CNN parameters and huge training cost. The recognition accuracy of the CNN network was 98.4% on the Plant Village data set of 39 different crops, and 98.7% on 10 tomato diseases. Karthik et al. [6] proposed attention embedded residual CNN, which effectively improved the performance of the model by combining residual connection and attention concentration mechanism, the average accuracy of fivefold cross validation in identifying three tomato diseases and healthy tomato leaves reached 98%. Lu et al. [7] processed 10 rice disease images by scale normalization, random sampling and mean normalization, and extracted feature maps by PCA and whitening, then sent them to CNN for training and testing, the accuracy of the test reached 95%. Li et al. [8] and Kumar et al. [9] proposed a crop recognition algorithm based on convolution neural network and support vector machine. Convolution neural network is used to extract high-level features of crop disease image, and then support vector machine is used for classification. In this way, the parameters of the model are reduced and the accuracy of the model is improved.

The research of the above scholars provides good ideas for the identification and classification of diseases and pests of crops. However, considering that agricultural diseases and pests generally occur in the field and the background noise in images is complex, it is difficult to directly apply the data set using only laboratory conditions to direct model training. Thus, further adjustments are needed for the models used.

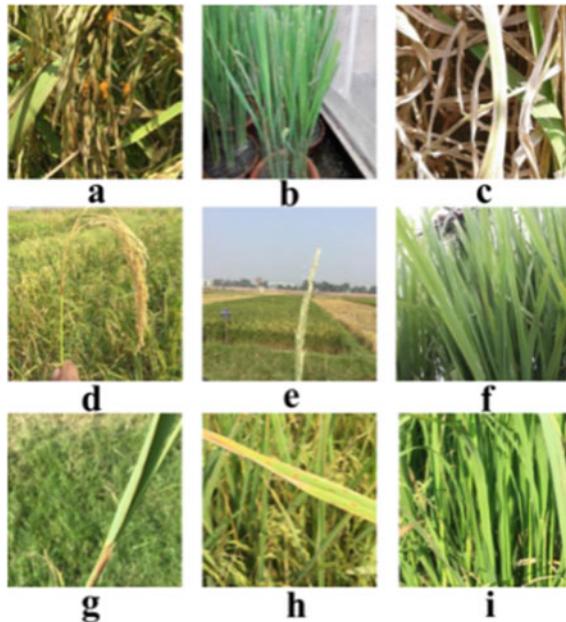
## 3 Materials and Methods

### 3.1 Experimental Environment

The experimental framework of this paper is Tensorflow2.0, the hardware environment is Google Cloud Collaboration, and the programming language is Python3.6.

### 3.2 Data Acquisition

The data in this paper are from a total of 1426 rice disease images collected by Rahman et al. [10] in Bangladesh Rice Research Institute (BRRI) from December 2017 to June 2018. The sample images are shown in Fig. 1. As shown in Table 1, the



**Fig. 1** Schematic diagram of rice diseases and insect pests (**a** is False Smut (disease). **b** is Brown Plant Hopper (pest). **c** is Bacterial Leaf Blight (disease). **d** is Neck Blast (disease). **e** is Stemborer (pest). **f** is Hispa (pest). **g** is Sheath Blight and/or Sheath Rot

**Table 1** Rice disease and pest data set for model training and model validation

Class name	Training set	Validation set
False smut	75	18
Brown plant hopper	57	14
Bacterial leaf blight	111	27
Neck blast	229	57
Stemborer	161	40
Hispa	58	15
Sheath blight and/or sheath rot	175	44
Brown spot	89	22
Health leaf	187	47
Total	1142	284

data set contains 9 different diseases and pests. In order to train the convolutional neural network model better, the data set is divided into training set and test set according to the ratio of 8:2.

### 3.3 Convolutional Neural Network Models

Convolutional neural network is essentially a forward neural network and deep learning method, which can optimize the network and improve the robustness through weight sharing, local connection and pooling [11]. With the continuous development in recent years, some convolutional neural network models are widely used in many fields because of their excellent feature extraction ability and robustness, Such as VGG16, InceptionV3 and ResNet50.

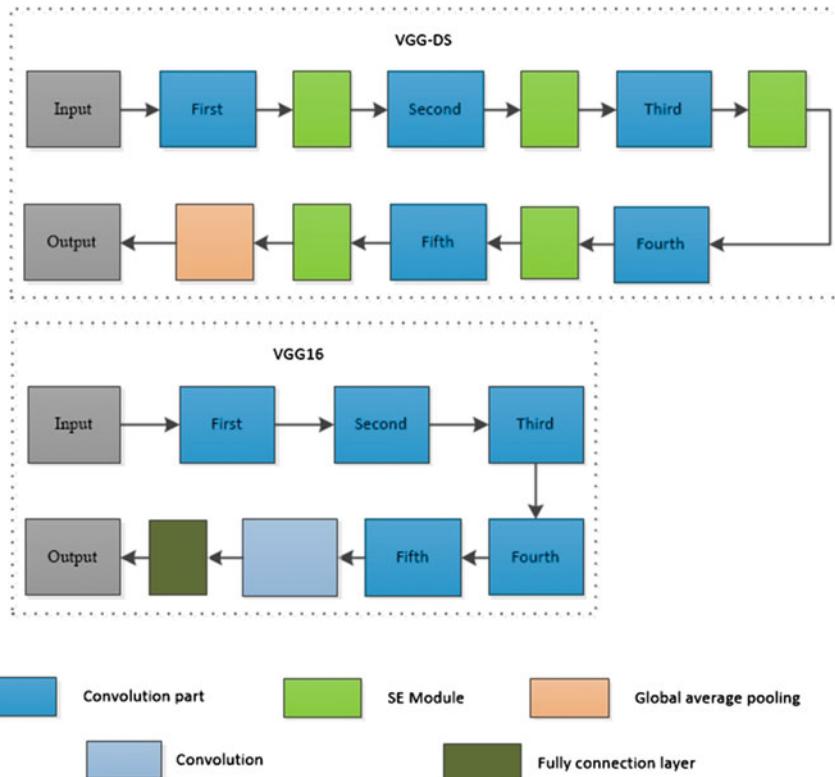
VGG16 was proposed by Simonyan et al. [12]. Compared with the earlier model AlexNet [13], VGG16 uses several  $3 \times 3$  small convolution kernels instead of large convolution kernels such as  $5 \times 5$ ,  $7 \times 7$ , so as to increase the network depth and ensure the learning features with fewer parameters by using multi-layer linear layer. InceptionV3 was proposed by Szegedy et al. [14]. Its main contribution is to further use small convolution kernel to replace large convolution kernel, and combine convolution layers with different convolution kernel size and step size to form Inception block, so as to further improve the ability of model feature extraction. ResNet50 was proposed by He et al. [15]. Its main contribution is to propose the residual structure, which enables the information from the previous residual block to flow smoothly to the next residual block through skip-connection and other techniques, and effectively avoids the gradient dispersion, explosion and network degradation because the network structure is too deep.

These three models are widely used in the research field because of their excellent performance. But in the real scene, it is very difficult to put these models into practical application due to the limitation of the model parameters on the application scenario and the limitation of the data set. Therefore, we improve VGG16 in the above three models and builds a lightweight network to solve the problems of large model parameters and small sample data set training.

## 4 VGG-DS (VGG-Depth-Separable Convolution and SE Module)

### 4.1 The Overview of VGG-DS

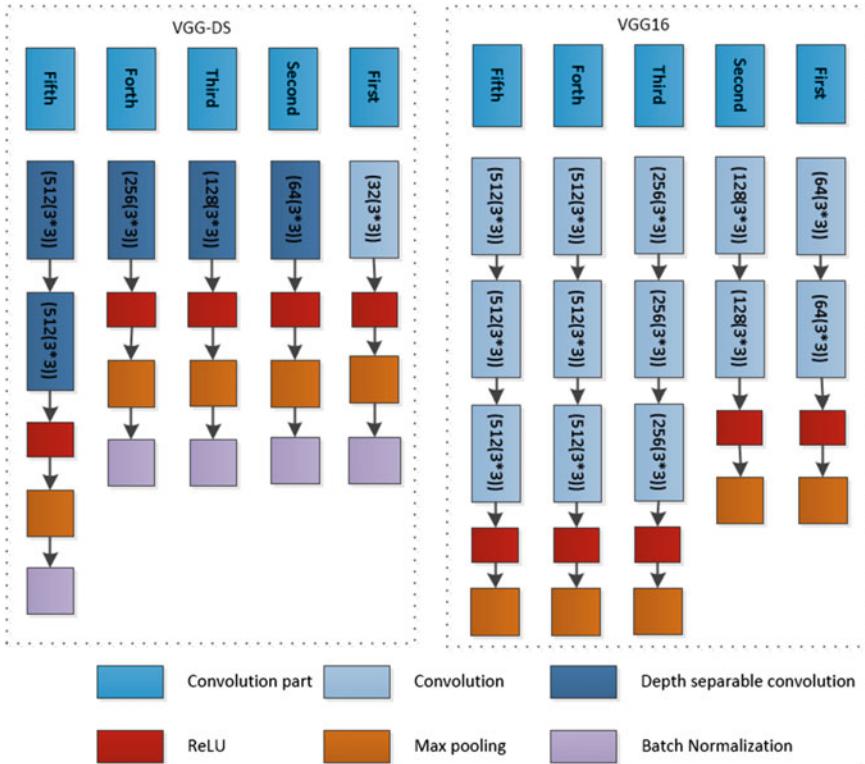
The problem of rice disease and insect classification and identification training is urgently needed, so this paper is based on VGG16, and establishes a new light quantitative network model, which we named VGG-DS.



**Fig. 2** Model structure comparison diagram of VGG-DS and VGG16

The contrast diagrams of structure between VGG16 and VGG-DS are shown in Figs. 2 and 3 respectively. The changes are described as follows:

- (1) To avoid the problem that VGG16 has too many parameters, VGG-DS has reduced the number of model channels.
- (2) To avoid the possible problem of over-fitting, VGG-DS adds batch normalization to each of the convolution parts.
- (3) The conventional convolution contains a lot of parameters. To reduce it, VGG-DS introduces deep separable convolution, which is designed to replace most of the conventional convolution.
- (4) In general, the lightweight networks' extraction ability of the characteristics is decreased compared to the traditional convolution neural network because of their model volume, therefore, VGG-DS introduces the SE module, so that the ability to extract features of the model can be improved. The workflow course is shown in the following text.



**Fig. 3** The structure of the convolution part

## 4.2 VGG-DS's Workflow

As shown in Figs. 2 and 3, VGG-DS's workflow is as follows:

- (1) VGG-DS first inputs the image data with the size of 224\*224\*3, and the data enters the conventional convolution layer. The channel number is 32 and the kernel size is 3\*3.
- (2) The data are secondly processed by Max Pooling and Batch Normalization respectively. And the data after the first convolution processing enters the deep separable convolution layer, and the channel number is 64, and the kernel size is 3\*3.
- (3) The data are thirdly processed by Max pooling and batch normalization respectively. The data processed by the first deep separable convolution goes into the second deep separable convolution operation. The number of channels is 128 and the kernel size is 3\*3.
- (4) The data are fourthly processed by Max Pooling and Batch Normalization respectively. The data processed by the second deep separable convolution goes

into the third deep separable convolution operation. The number of channels is 256 and the kernel size is 3\*3.

- (5) The data fifthly enters the last two depth separable convolution layers. The number of channels is 512, and the kernel size is 3\*3.
- (6) Finally, the Max Pooling and Batch Normalization are performed respectively. Then, the data information obtained from the above processing is global averaged pooling, and output in the full connection layer.

### 4.3 Batch Normalization

Batch normalization is an effective data regularization method [16]. The basic idea is to force the neural network of each layer of input neurons to return to the standard normal distribution with the mean value of 0 and the variance of 1. This not only accelerates the convergence of the model and destroys the original data distribution, but also alleviates the over fitting phenomenon to a certain extent. The equation is as follows:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

$$\delta = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (2)$$

$$\bar{x}_i = \frac{x_i - \mu}{\sqrt{\delta^2 + \epsilon}} \quad (3)$$

where  $x_i$  is the  $i$ th pixel value of the data set.  $n$  is the total number of pixels.  $\mu$  is the mean value of  $n$  pixels.  $\delta$  is the variance.  $\bar{x}_i$  is the normalized pixel value.  $\epsilon$  is a minimal positive value to ensure that the denominator in Eq. (3) is greater than 0. The workflow of batch normalization is as follows:

- (1) The mean value of elements in mini batch is obtained by using Eq. (1).
- (2) The variance of mini batch is solved by using Eq. (2).
- (3) Each element is normalized by using Eq. (3).
- (4) The scale-scale transformation is used to transform the transformation into the original distribution, so as to realize the identity transformation and improve the nonlinear expression ability of the step size network.

In the convolution neural network, the batch normalization is added to make the optimization space of the model smoother in the training stage, and improve the training accuracy and efficiency of the model.

#### 4.4 Depth Separable Convolution

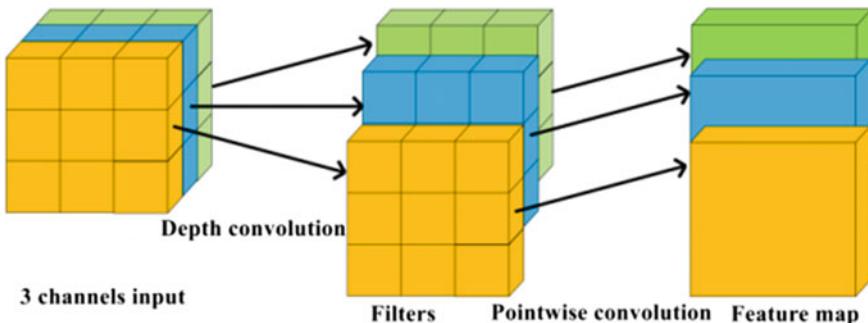
As a variant of conventional convolution, deep separable convolution is widely used in lightweight networks because it can greatly reduce the model parameters [17]. As shown in Fig. 4, deep separable convolution mainly includes two parts, namely, depth-wise convolution and point-wise convolution. The basic idea is to split the feature learning from the standard convolution operation through the spatial feature learning step and the channel combination step. Suppose an image with a  $a \times a$  pixel and three channels passes through a conventional convolution layer with convolution kernel size of  $b \times b$  and output channel number of 4. There are four filters in the convolution layer, and each filter contains three convolution kernels with each convolution kernel size of  $3 \times 3$ . The volume of parameters is:

$$N_{\text{std}} = 4 * 3 * 3 * 3 = 108 \quad (4)$$

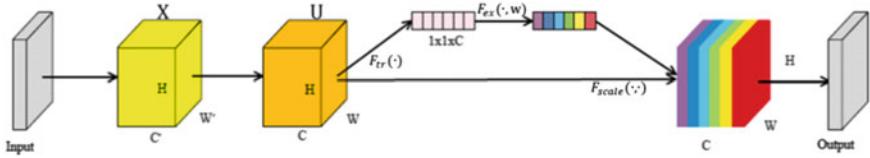
When the same image is input into the depth separable convolution, the first convolution operation is performed, and the depth separable convolution is completely carried out in the two-dimensional plane. The number of kernel sizes is the same as the number of channels in the upper layer, so the image will only pass through three filters, and each filter contains a kernel size with the size of  $3 \times 3$ . Therefore, the number of parameters is:

$$N_{\text{depthwise}} = 3 * 3 * 3 = 27 \quad (5)$$

It can be seen that the depth separable convolution can greatly reduce the model parameters, which is helpful for the reduction of model size.



**Fig. 4** Diagram of depth separable convolution



**Fig. 5** SE module structure diagram

#### 4.5 SE Module

SE module comes from SENet [18], which is one of the representatives of lightweight network that has been developing in recent years. Compared with deep separable convolution, SENet pays more attention to the relationship between the channel numbers of feature graph, and hopes that the model can automatically learn the importance of different channel features. Based on this, the SE module derived from SENet mainly includes two operations, Squeeze and Excitation.

As shown in Fig. 5, the SE module first performs Squeeze operation on the convoluted feature graph to obtain the global feature of the channel. The equation is as follows:

$$Z_c = F_{sq}(u_c) = \frac{1}{H * W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j), \quad z \in R^c \quad (6)$$

After the Squeeze operation, we need to carry out the Excitation operation on the global features, learn the relationship between each channel, and get the weight of different channels, and finally multiply the original feature graph to get the final feature. The equation is as follows:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \text{ReLU}(W_1 z)) \quad (7)$$

In addition, SE module has good compatibility with other models, which can effectively improve the feature extraction ability of the model. Therefore, the VGG-DS proposed in this paper is added to this module to further improve the training accuracy of the model.

## 5 Results

For confirming the effect of VGG-DS, four large-scale CNN models such as VGG16, InceptionV3, ResNet50 and ResNet101, two three lightweight network models MobileNet and, Xception, and the simple CNN built by Rahman et al. [10] are introduced. Reduce the influence of irrelevant factor, the hyperparameters of the

**Table 2** Model training hyperparameters

Hyperparameter	Number/type
Learning rate	0.01
Epoch	50
Learning rate attenuation coefficient	0.8
Optimizer	Adam
Batch size	32

**Table 3** Comparison of the number of model parameters, model training time and model accuracy

Models	Parameters	Elapsed time/s	Accuracy/%
VGG-DS	826,753	370	93.66
Simple CNN	0.8 M		94.33
MobileNet	3,238,089	503	89.42
Xception	20,879,921	2306	91.12
VGG16	134,297,417	910	20.7
InceptionV3	21,821,225	1012	82.07
ResNet50	23,606,153	755	88.98
ResNet101	42,676,617	1150	81.83

model participating in the experiment are unified, and the specific values are shown in Table 2.

After unifying the hyperparameters, we use 5-folds cross validation to train all the above models. The training results are shown in Table 3. In terms of model parameters, the parameters of VGG-DS and simple CNN are about 0.8 M, which is only 0.6% compared with VGG16, which has the largest number of parameters, in which VGG-DS only consumes 370 s for model training, which is nearly 8 times different from Xception, which has the longest time. From the perspective of model accuracy, due to the small number of model parameters and relatively simple structure, the lightweight network model needs fewer data features for model training. Therefore, when training small sample data sets, the lightweight network model has more advantages, and the model accuracy is generally higher than large-scale models. Among them, accuracy of VGG16 is only 20.6% because of over-fitting caused by large parameters, while the accuracy of VGG-DS is 93.66%, which is better than large-scale network model and traditional lightweight network model. But it still has a small gap compared with simple CNN. Therefore, we consider from the perspective of hyperparameters, hoping to achieve the optimal model of VGG-DS by adjusting the hyperparameters.

For improve the effect of VGG-DS, we make different attempts on the optimizer and learning rate. The specific results are shown in Table 4, when the optimizer is Adamax and the learning rate is 0.01, the accuracy is the highest, which reaches 95.09%. The optimizer and learning rate have a great influence on the accuracy.

**Table 4** Accuracy comparison of VGG-DS model under different optimizers and learning rates

Optimizer/learning rate	Accuracy		
	0.01	0.001	0.0001
Adam	93.66%	94.93%	87.66%
SGD	53.65%	27.96%	27.32%
RMSprop	91.90%	94.19%	83.5%
Adamax	<b>95.09%</b>	90.70%	59.15%
Nadam	93.24%	94.58%	86.52%
Adagrad	83.24%	29.27%	27.80%
Adadelta	34.02%	26.34%	15.12%

## 6 Conclusion and Prospect

In order to solve the problem of large-scale network difficult to be applied in mobile terminals caused by too large parameters and too small data set in the process of using CNN to identify and classify rice diseases and pests. A new lightweight network VGG-DS is built with depth-separable convolution and SE module. The experiment result shows that the parameters of VGG-DS are only 826,753, and the accuracy is 93.66%. After adjusting the optimizer and learning rate, the accuracy of VGG-DS is further improved, up to 95.09%. The next work is to adapt the model of the mobile terminal and optimize this model to improve its robustness.

**Acknowledgements** This work was supported in part by the Key Realm Research and Development Program of Guangdong Province under Grant 2019B020214002, in part by the Research Project of Beijing Municipal Social Science Foundation under Grant 20GLB026, in part by the Beijing Wuzi University 2020 Education and Teaching Reform Project, in part by the National Natural Science Foundation of China under Grant 71771028, in part by the Beijing Key Laboratory of Intelligent Logistics System under Grant BZ0211, and in part by the Beijing Intelligent Logistics System Collaborative Innovation Center.

## References

1. Pierce, F.J., Nowak, P.: Aspects of precision agriculture. *Adv. Agron.* **67**, 1–85 (1999)
2. Peng, W., Yubin, L., Xuejun, Y., Ziyao, C., Linhui, W., Zhenzhao, C., Lu, Y., Jinbao, H.: Research on weed identification in paddy field based on deep convolutional neural network. *J. South China Agric. Univ.* **41**(06), 75–81 (2020)
3. FAO Crops production accessed [Online], <http://www.fao.org/faostat>. Last accessed 2020
4. Xihai, Z., Yue, Q., Fanfeng, M., et al.: Identification of maize leaf diseases using improved deep convolutional neural networks. *IEEE Access* **6**, 30370–30377 (2018)
5. Agarwal, M., Gupta, S.K., Biswas, K.K.: Development of efficient CNN model for tomato crop disease identification. *Sustain. Comput. Inform. Syst.* **28**, 100407 (2020)
6. Karthik, R., Hariharan, M., Anand, S., et al.: Attention embedded residual CNN for disease detection in tomato leaves. *Appl. Soft Comput.* **86**, 105933 (2020)

7. Yang, L., Shujuan, Y., Nianyin, Z., et al.: Identification of rice diseases using deep convolutional neural networks. *Neurocomputing* **267**, 378–384 (2017)
8. Yang, L., Jing, N., Xuewei, C.: Do we really need deep CNN for plant diseases identification? *Computers Electron. Agric.* **178**, 105803 (2020)
9. Prabira, K.S., Nalini, K.B., Amiya, K.R., et al.: Deep feature based rice leaf disease identification using support vector machine. *Computers Electron. Agric.* **175**, 105527 (2020)
10. Chowdhury, R.R., Preetom, S.A., Mohammed, E.A., et al.: Identification and recognition of rice diseases and pests using convolutional neural networks. *Biosys. Eng.* **194**, 112–120 (2020)
11. Xiangwu, D., et al.: Weed identification in rice field based on convolutional neural network and transfer learning. *J. Agric. Mech. Res.* **43**(10), 167–171 (2021)
12. Karen, S., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
13. Alex, K., Ilya, S., Geoffrey., E.H.: Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* **25**, 1097–1105 (2012)
14. Christian, S., Vincent, V., Sergey, I., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
15. Kaiming, H., Xiangyu, Z., Shaoqin, R., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE (2016)
16. Sergey, I., Christian., S.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, vol. 37, pp. 448–456. PMLR (2015)
17. Jianbo, G., Yuxi, L., Weiyao, L., et al. Network decoupling: from regular to depthwise separable convolutions. arXiv preprint [arXiv:1808.05517](https://arxiv.org/abs/1808.05517) (2018)
18. Jie, H., Li, S., Samuel, A., et al.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

# Compressed Channel Attention Mechanism for 3D Medical Image Segmentation of Liver



Yuwei Liao , Lianglun Cheng , and Weida Lin

**Abstract** More and more Convolutional Neural Networks (CNNs) are used in computer vision, especially in the field of medical images. Since most of the medical data in clinical practice are three-dimensional, which is mainly obtained from imaging techniques such as MRI and CT, the use of the previous two-dimensional neural network becomes untimely. In this work, an end-to-end trained 3D image segmentation network combined with the Compressed Channel Attention Module (CCAM) is proposed to learn to predict the segmentation of the entire liver at one time. We embed the CCAM in the network structure of up-sampling and down-sampling. First, we divide the feature map obtained by down-sampling into two parts. The first part performs global average pooling and maximum pooling and then splicing, and the other part performs  $1 \times 1$  convolution on the feature map. Then the two parts are merged, and finally spliced with the up-sampled feature map to realize the use of down-sampling features to monitor the up-sampled features, focusing on specific livers, and suppressing irrelevant areas in the input image. Experiments have been carried on the LITS dataset. Compared with the existing segmentation methods, the proposed method has better segmentation performance in both subjective and objective evaluation.

**Keywords** 3D image segmentation · Attention mechanism · Deep learning

## 1 Introduction

Segmentation is one of the most critical and challenging tasks in medical image analysis. It plays a vital role in the detection of lesions, diagnosis of diseases and formulation of protocols. The purpose of medical image segmentation is to delineate the interesting anatomical structures, such as tumors, organs, and tissues, in a semi-automatic or fully automatic way, which is conducive to the diagnosis of patients' conditions and the provision of complementary treatment.

---

Y. Liao · L. Cheng · W. Lin  
Guangdong University of Technology, Guangzhou 510006, China  
e-mail: [yuwei333777@163.com](mailto:yuwei333777@163.com)

Over the past few years, more and more international segmentation challenges have emerged, which provides a transparent platform for us to fairly and equitably evaluate different approaches. Currently, popular medical image segmentation tasks mainly include brain and brain tumor segmentation [1, 2], lung segmentation and tuberculosis [3, 4], breast cancer segmentation [5, 6], liver segmentation [7, 8], and so on. Thanks to the improvement of computer computing power, researchers are no longer limited to the 2D dataset of medical images, and more and more people devote themselves to the research of 3D medical images. In particular, since diagnostic images are usually in a three-dimensional format, being able to test segmentations by feeding the entire volume into the network, has a particular convenience. It has changed the cumbersome pattern that doctors used to process original 3D image slices [9–11] into 2D, and the directly predicted three-dimensional organs or lesions are more vivid and straightforward.

In recent years, various variants [12–14] based on U-Net [15] have emerged in an endless stream. Using a 3D network directly can better mine the semantic features of 3D data. Inspired by this, Cicek et al. [16] replaced all 2D convolution operations with 3D convolution operations, and the network was basically consistent with the previous U-Net architecture of Ronneberger et al. [15] and used elastic deformation for effective data enhancement during training. Although it largely solves the awkward situation that 3D images are sent into the model one by one for training and greatly improves the training efficiency, 3D U-Net only contains three layers of down-sampling due to the huge training parameters, which is insufficient to extract deeper semantic information. Afterward, Milletari et al. [17] added an additional layer of down-sampling on the basis of the 3D U-Net [16] through the residual connections, which can accelerate the network convergence and avoid falling into the disappearance of the gradient. The Voxresnet proposed by Yu et al. [18] also uses the residual connection, while using deep supervision to change the loss function into a fusion of multi-layer output. However, the common point of these 3D networks is that they do not pay attention to the spatial information of the image, and it is difficult to distinguish organs or tissues that are close to the target area, and their parameters are extensive.

In this work, we chose the LiTS [7] dataset as our segmentation task, with the goal of segmenting the volume of liver CT. In practical application, as the abdomen is one of the parts with the most organs in the human body, some adjacent tissues and organs are close to the liver in morphology, so how to separate the liver from them has become the difficulty of this task. In this paper, a 3D network-based Compressed Channel Attention Module (CCAM) is proposed. The method monitors the up-sampled features through the down-sampled features and focuses attention on a specific liver to suppress the irrelevant regions in the input image. The qualitative and quantitative experimental results prove the superior segmentation performance of the proposed method compared to exiting classic schemes.

## 2 Related Work

In recent years, many methods for liver segmentation have been proposed, especially the deep learning-based segmentation method.

Li et al. [19] proposed a novel H-DenseUNet for liver segmentation, which cascaded features obtained from a 2D network with 3D raw data into a 3D network for training. Although the network is a good solution to the lack of volume context information in 2D training and the high computational memory consumption of 3D training, it is still not a good way to distinguish between liver and its surrounding tissues.

Since most 3D networks are very complex, in order to reduce the number of parameters in the network model and the dependence on memory capacity, many researchers use depth separability to apply depth separable convolution to 3D networks. Lei et al. [20] proposed a lightweight V-Net (LV-Net) for liver segmentation. Generally, the depth separable convolution is divided into channel-wise convolution and point-wise convolution, but due to the fragmented calculation process, its efficiency in the existing convolutional neural network implementations is not high enough, and the iteration speed during training may be plodding.

In response to these problems, we propose a novel Compressed Channel Attention Module (CCAM) network to focus on the target area entirely. This method uses V-Net [17] as the baseline, extracts deep semantic features through the down-sampling stage, and then stitches the low-dimensional and high-dimensional features through the skip-connection embedded CCAM in the up-sampling stage.

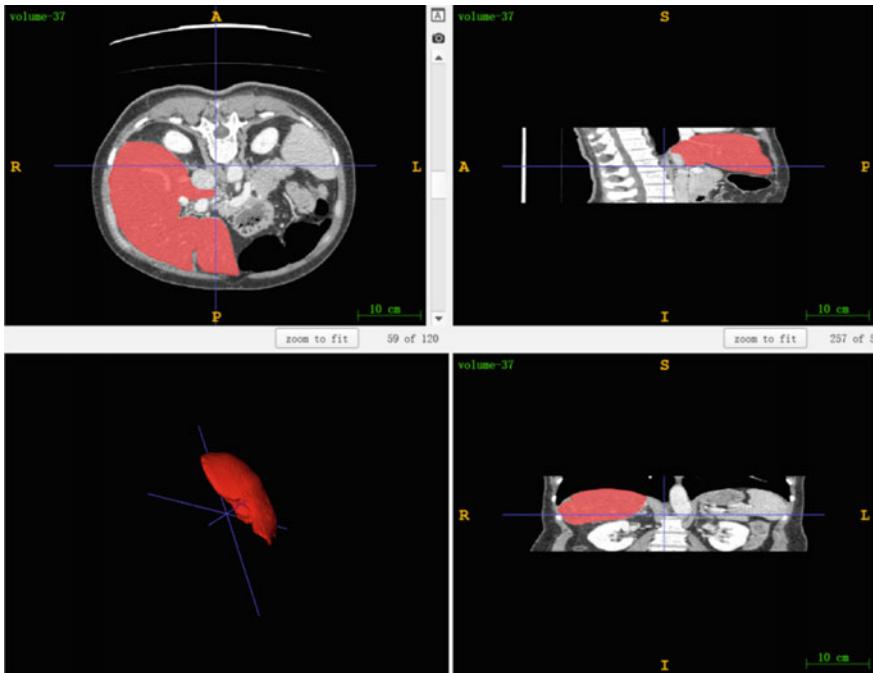
## 3 Method

Our method is mainly divided into three steps, as shown below.

### 3.1 *Image Preprocessing*

First, we will convert the CT value to the standard HU value, enhancing the contrast between the different organs. The HU value is device-independent, and values in different ranges can represent different organs. Then, we use a commonly used gray transformation method—histogram equalization, the gray value outside the threshold is truncated. The equalization is helpful to the extension of the image histogram, the gray level range of the image after equalization is wider, and the image's contrast is enhanced effectively. Finally, all sections of the abdomen were normalized, and only those containing the liver were extracted, and the rest were excluded.

This is schematically represented in Fig. 1.

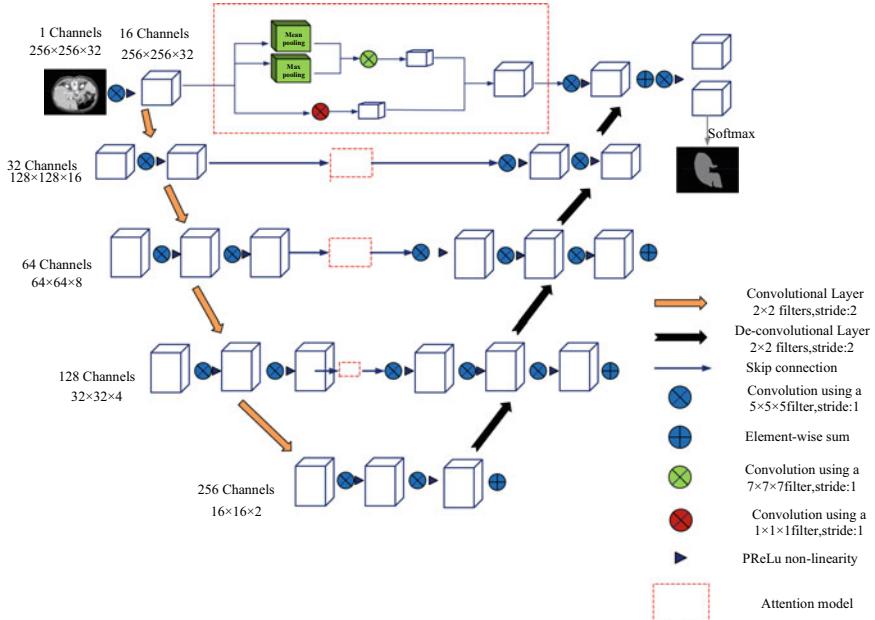


**Fig. 1** CT images of the abdomen labeled with liver and preprocessed by us. This data is part of the LiTS2017 challenge dataset

### 3.2 Architecture of Network

In this section, the Compressed Channel Attention Module (CCAM) V-Net network architecture is presented (illustrated in Fig. 2). The network is mainly divided into two parts, down-sampling and up-sampling, which makes the network look like a V shape. On the left side of the network, we obtain the deep semantic information of the image through four steps of down-sampling. Each step is achieved by convolving with  $2 \times 2 \times 2$  voxel-wide kernels and applying a step size of 2. At the same time, in each stage, we use  $5 \times 5 \times 5$  voxel-wide kernels to perform 1 to 3 convolution operations to make it learn the Residual Function. On the right side of the network, we use transposed convolution to achieve up-sampling, and use skip connection to overlay the feature maps of the previous layer through the CCAM. In this way, we can collect fine-grained details that will be lost in down-sampling, and we can improve the quality of the final contour prediction.

Our Compressed Channel Attention Module (CCAM) is creatively embedded in the network. First, the CCAM divides the feature map obtained by down-sampling into two parts. The first part performs global average pooling and maximum pooling and then splicing. In this way, the network not only retains the global features, but also pays attention to specific features. The other part performs  $1 \times 1$  convolution



**Fig. 2** Schematic diagram of our network structure. We use Pytorch to customize our volume convolution to process 3D data. The dotted rectangle in the figure represents the Compressed Channel Attention Module (CCAM) we designed, which will be used by all skip connections and have the same content

on the feature maps, reducing the number of network parameters. Then the two parts are merged and finally spliced with the up-sampling feature map to realize the use of down-sampling features to monitor up-sampling features, focusing on specific livers and suppressing irrelevant areas in the input image.

### 3.3 Training and Testing

Due to the large size of 3D images, the input of the whole image into the network requires a large amount of GPU video memory. Therefore, we adopted the patch-based training method and randomly selected fixed patches for each training set. In order to avoid that the randomly selected patch does not contain the liver or only a tiny part of the liver area, we need to choose a larger patch size as much as possible. Ultimately, based on our existing experimental equipment, all volumes processed by the network have a fixed size of  $256 \times 256 \times 32$  voxels and a spatial resolution of  $1 \times 1 \times 1$  mm. In the test phase, the sliding patch method is used, the size is the same as the training, and finally, each patch is predicted and spliced back.

For the details of the training phase, we set the training epoch to 50, and the batch size is 2. Since annotated medical images are not easy to obtain, one or more experts with professional medical knowledge are required to describe the shape of the organ or lesion manually. Therefore, before reading in the data, we performed a flip, elastic deformation and other data enhancements [20, 21] to improve the generalization ability of the model. In theory, we use BCEWithLogitsLoss. This loss combines the Sigmoid layer and BCELoss (Binary Cross-Entropy) into one formula. This version is more stable than separate use. By combining operations into one layer, we use log-sum-exp techniques to achieve numerical stability.

$$L_{BCE} = -\omega_n [y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))] \quad (1)$$

## 4 Experimental Results and Analysis

We will begin with the following two chapters.

### 4.1 Experimental Environment and Evaluation Metric

We train our network on 99 abdominal CT volumes and test the effect of the network on 31 abdominal CT volumes and use Adam as our network optimizer which the learning rate is set to 0.002. The experimental equipment includes two 2080Ti graphics cards, 32 GB of memory and a CPU model of i9-10900X. See Sect. 3.2 for other experimental details. See Sect. 3.2 for other experimental details.

In order to quantitatively demonstrate the superiority of our segmentation method, we use four popular segmentation quality metrics: Dice (Dice Similarity Coefficient), Jaccard (Jaccard similarity coefficient), Precision, and Recall.

Dice is one of the most popular performance measures for medical image segmentation. Similar to IOU, this measure is used to calculate the similarity between two samples, which is essentially a measure of the overlap between the predicted value and the real value. The range of values is [0,1]. The best segmentation result is 1 and the worst is 0. The calculation formula is defined as

$$Dice(A, B) = \frac{2 \times |A \cap B|}{A + B} \times 100\% \quad (2)$$

where A is a predicted segmentation result and B is a real segmentation result. Another formula is defined as

$$Dice = \frac{2TP}{FP + 2TP + FN} \quad (3)$$

where TP means true positive, it is judged as a positive sample, in fact, it is also a positive sample. FP means false positive, which is judged as a positive sample but is a negative sample. FN means false negative, it is judged as a negative sample, but in fact, it is a positive sample.

The second evaluation metric is Jaccard. Given two sets A and B, the Jaccard coefficient is defined as the ratio of the size of the intersection of A and B to the size of the union of A and B, defined as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (4)$$

The other two evaluation metrics are Precision and Recall. The accuracy rate is based on our prediction results, and it indicates how many of the samples whose predictions are positive are truly positive samples. The recall rate is for our original sample. It indicates how many positive examples in the sample are predicted correctly.

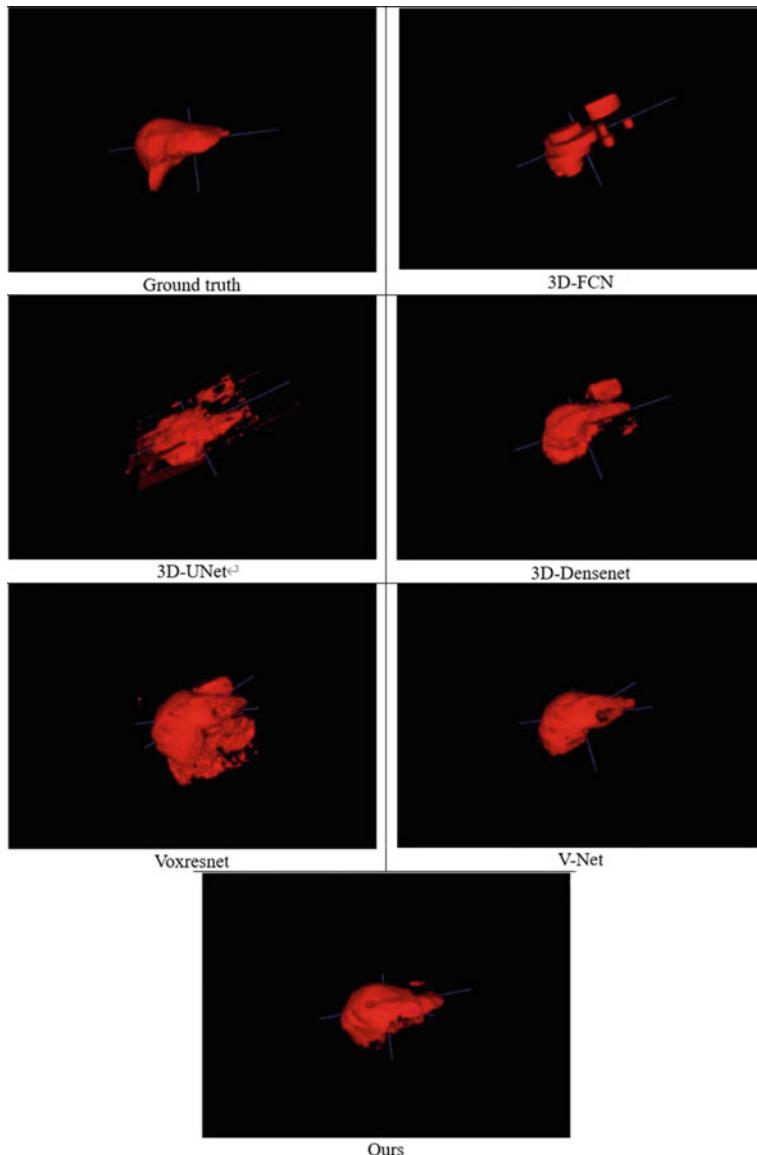
## 4.2 Results and Analysis

Five classic segmentation methods will be used to compare with our method, including 3D-FCN [22], 3D-UNet [16], 3D-Densenet [23], VoxResNet [18], and V-Net [17]. Except for the difference in network structure, all training methods and testing methods are the same, and various parameters are also the same. In order to intuitively reflect the effect of segmentation, we selected a sample in the test set to show the segmentation effect of different methods, as shown in Fig. 3. At the same time, we summarized the average evaluation metrics of the 31 test set samples in Table 1. The best value of the quality index is shown in bold.

As we can seen from Fig. 3, both our method and VNET can basically segment the liver completely, but other methods are only just passable. Furthermore, by analyzing the results of each test set, we find that the overall performance of VNet is good, but the performance on individual volumes is extremely poor, while our method performs well, so the overall level is improved (as shown in Table 1). In a word, our method shows its effectiveness.

## 5 Conclusion

In this paper, we propose a novel Compressed Channel Attention Module (CCAM) for 3D medical image segmentation. Through this module, the down-sampling feature is used to monitor the up-sampling feature, focus on the specific liver and suppress the irrelevant areas in the input image. In the future, we will study the performance of our attention module in more segmentation competitions, and constantly improve it.



**Fig. 3** Comparison diagram of experimental results

**Table 1** Comparison between classic segmentation networks and ours

Methods	Metrics			
	Avg. dice	Avg. precision	Avg. jaccard	Avg. recall
3D FCN [22]	0.533	0.704	0.402	0.447
3D UNet [16]	0.675	0.747	0.560	0.628
Densenet [23]	0.622	0.747	0.495	0.567
VoxResNet [18]	0.602	0.477	0.462	<b>0.843</b>
V-Net [17]	0.605	0.749	0.525	0.572
Ours	<b>0.719</b>	<b>0.770</b>	<b>0.615</b>	0.707

## References

1. Brain Tumor Segmentation (BraTS) Challenge 2020, [www.braintumorsegmentation.org/](http://www.braintumorsegmentation.org/). Last accessed 18 May 2020
2. Martino, A.D.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry **19**(6), 659–667 (2013)
3. Armato, S.G.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Acad. Radiol. **14**(12), 1455–1463 (2007)
4. COVID-19CT, <https://www.kaggle.com/andrewmvd/covid19-ctscans/tasks?taskId=811>. Last accessed 14 May 2020
5. Shen, L., Margolies, L.R., Rothstein, J.H., et al.: Deep learning to improve breast cancer early detection on screening mammography. Sci. Rep. 12495 (2019)
6. Lee, R.S., Gimenez, F., Hoogi, A., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data **4**, 170177 (2017)
7. Bilic, P., Christ, P.F., Vorontsov, E., et al.: The liver tumor segmentation benchmark (LiTS) (2019)
8. SLIVER07, <https://sliver07.grand-challenge.org/>. Last accessed Feb 2019
9. Chen, H., Yu, L., Dou, Q., Shi, L.: Automatic detection of cerebral microbleeds via deep learning based 3d feature representation. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 764–767 (2015)
10. Prasoon, A., Petersen, K., Igel, C., Lauze, F.: Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 246–253. Springer (2013)
11. Roth, H. R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J.: A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 520–527. Springer, Berlin (2014)
12. Alom, M.Z., Hasan, M., Yakopcic, C., et al.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv 1802.06955 (2018)
13. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., et al.: Unet++: a nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer, Cham (2018)
14. Xiao, X., Lian, S., Luo, Z., et al.: Weighted Res-UNet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education(ITME), pp. 327–331 (2018)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional Networks for Biomedical Image Segmentation, pp. 234–241. MICCAI Springer (2015)

16. Cicek, Ö., Abdulkadir, A., Lienkamp, S., et al.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. (MICCAI), pp. 424–432. Springer (2016)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
18. Chen, H., Dou, Q., Yu, L., Heng, P.A.: Voxresnet: deep voxel wise residual networks for volumetric brain segmentation. arXiv preprint [arXiv:1608.05895](https://arxiv.org/abs/1608.05895) (2016)
19. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE Trans. Med. Imag **37**(12), 2663–2674 (2018)
20. Kisantal, M., Wojna, Z., Murawski, J., et al.: Augmentation for small object detection. In: 9th International Conference on Advances in Computing and Information Technology (2019)
21. Ekin, D., Barret, Z., Dandelion, M., et al.: Auto augment: learning augmentation strategies from data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2019)
22. Li, B.: 3D Fully convolutional network for vehicle detection in point cloud (2016)
23. Bui, T.D., Shin, J., Moon, T.: 3D densely convolutional networks for volumetric segmentation (2017)

# Action Detection Based on Transfer Learning of Human Pose Estimation



Weida Lin<sup>ID</sup>, Zhuwei Wang<sup>ID</sup>, and Yuwei Liao<sup>ID</sup>

**Abstract** The purpose of action detection is to detect whether the specified action behavior occurs in the video, and find out the time when the action occurs. Predecessors introduced pose information into the field of action detection through RNN and GCN, but because of the gradient disappearance problem of recurrent neural network (RNN) and the problem of graph convolution (GCN) calculation efficiency, they encountered bottlenecks. In this paper, we propose to use convolutional neural network (CNN) to introduce pose information into behavior detection by means of transfer learning. We use CenterNet to obtain high-semantic information to improve the accuracy of behavior detection. In order to generate a more suitable pose heatmap of the application scenario, our CenterNet pose estimation network participates in the training, and constantly updates the parameters of the pose estimation network according to the new dataset. We conducted a large number of ablation experiments on the Ur fall detection dataset and RWF-2000 dataset. On the RWF-2000 data set, we added a pose estimation network to generate a pose heatmap and input it into the subsequent network, the accuracy increased from 79.0 to 83.75%.

**Keywords** Action detection · Transfer learning · Pose estimation

## 1 Introduction

Action detection aims to find out whether an action behavior occurs in the video, and find out the specific time when the action occurs. Action detection are generally used in video analysis systems to detect the occurrence of abnormal behaviors, which can reduce the labor intensity of manual video analysis. When the behavior detection reaches real-time, it becomes possible to automatically alarm in time after abnormal behavior is found, which will exert huge social and economic benefits.

---

W. Lin · Z. Wang (✉) · Y. Liao  
GuangDong University of Technology, Guangzhou 510006, China  
e-mail: [wangzhuowei0710@163.com](mailto:wangzhuowei0710@163.com)

Z. Wang  
Wuhan Donghu University, Hubei 430074, China

Action detection is mainly divided into offline action detection and online action detection. Offline action detection reads an entire video at once, and then locates the time when the action occurred in the entire video. Online action detection will continuously read in new frames and find abnormal behaviors in the video in time, so the prediction speed is required to be real-time. It can be seen that online action detection is more in line with realistic requirements, and this paper will focus on online action detection.

Generally, actions only occupy a small part of the video, so a large amount of interference needs to be dealt with in the video sequence. In the traditional method, iDT [1] extracts trajectories, further extracts feature descriptors such as HOF, HOG, and finally encodes the features, and then trains the SVM classifier for action recognition. TwoStream [2], TSN [3], etc. extract the spatial and temporal features of video data through two branches, and finally combine the features for action recognition. The pose estimation aims to predict the position of human joints. The joints of the human body contain a lot of action information. Li et al. proposed a joint behavior recognition method based on convolutional neural network. Li et al. proposed a action recognition method based on convolutional neural networks and joint positions.

Since most of the current action recognition datasets do not have the annotation information of human pose, the common method of the current action recognition algorithms based on skeleton is to use the pose estimation algorithm to predict the pose information and save it locally. However, the speed of the method saved locally cannot be guaranteed, and the pose information generated by the pose estimation model trained using other datasets on the new dataset cannot be guaranteed to be accurate. The other method uses a multi-task method to predict joint points and recognize behavior at the same time. However, this will bring a larger amount of parameters, and requires pose annotations in the training dataset.

So we propose to use the transfer learning method to introduce the heatmap of human joint points into the behavior recognition algorithm, which can extend the joint behavior recognition algorithm to all behavior recognition algorithms. Secondly, it can generate pose information that is more suitable for our application scenarios.

Therefore, the main contributions of this paper are as follows:

- (1) Try to use the pose estimation heatmap with high semantic information as the input of the neural network, and obtain faster convergence speed and accuracy.
- (2) Use the model parameters of the pose estimation in CenterNet for transfer learning, and participate in the training and learning of the pose estimation network parameters in CenterNet, so that the generated pose heatmap is more suitable for our algorithm and gains stronger robustness.
- (3) Our accuracy on the RWF-2000 dataset has increased from 79.0 to 83.75%, on the UR-fall detection dataset the accuracy on the validation set has increased from 91.7 to 100.0%, and the accuracy of the test set has increased from 75.0 to 83.3%.

## 2 Related Work

### 2.1 Action Detection

In the context of deep learning, researchers usually edit videos so that they can directly perform end-to-end feature extraction and recognition on the video. Therefore, action detection usually uses a sliding window to split the video segment, and then send it to the action recognition model for prediction. Simonyan et al. [2] proposed to use two-stream CNN to capture time sequence information such as dense optical flow of the image and image spatial position information, and finally directly fusion and classification of the features of the two branches. Since TwoStream [2] only operates one video segment when capturing timing information, Wang et al. [3] proposed to divide the entire video into K video segments, and each video is processed separately and integrated to predict that the entire video belongs to each The probability of the action category.

Tran et al. [5] used deep 3D convolutional networks for behavior recognition, and proposed that 3D convolutional networks are more suitable for temporal and spatial feature extraction than 2D networks. Feichtenhofer et al. [6] used different frame rate sampling to capture sparse frame image video clips, and then used 3D convolution for feature extraction.

Shahroudy et al. [7] proposed to use RNN to simulate the long-term correlation of human skeleton for action recognition. Yan et al. [8] proposed a spatiotemporal graph convolutional network based on GCN for behavior recognition. They used the human body pose information with high semantic information and obtained a huge improvement. However, as the network deepens, the RNN-based model tends to disappear in the training stage and becomes difficult to train. The GCN model introduces prior knowledge of the human body topology, but most of the current GCN algorithms are not well optimized, resulting in inefficient calculation. The current method based on CNN, its performance is not inferior to GCN, and the research and optimization work in recent years has continuously improved the computational efficiency of CNN. Therefore, this paper will model the pose information based on the CNN model.

### 2.2 Transfer Learning

In recent years, in order to solve the problem of insufficient training data for deep learning, researchers have transferred knowledge from other source domains to the target domain, usually called transfer learning. At present, almost all deep learning uses a powerful backbone network as pre-training, such as RestNet [9], DLA [10], VGG [11], InceptionNet [12, 13] and so on. For example, faster-rcnn [14] in the field of target detection uses ResNet's pre-trained model as the backbone, and yolov3 [15] uses Darknet-53 [16].

In this paper, we will use the knowledge learned by the pose estimation network in CenterNet [17] to extract human pose information as the subsequent training data. And participate in the training of this part of the network parameters to generate a heatmap that is more suitable for the new dataset.

### 3 Method

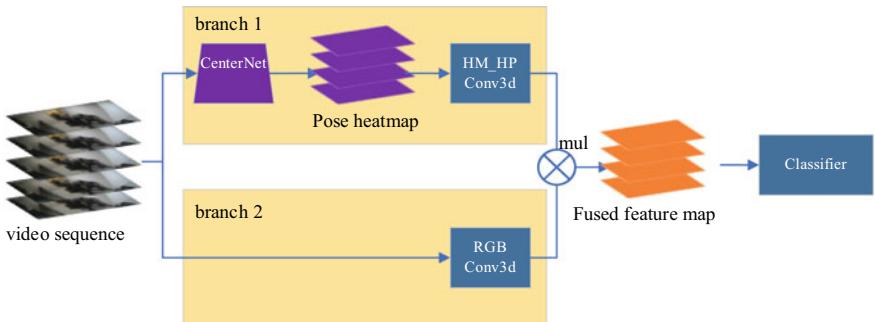
For each video segment, we divide it into two branches to predict, as shown in Fig. 1. In the first branch, we use CenterNet to generate a heatmap  $\hat{\Phi} \in \mathfrak{R}^{\frac{w}{R} \times \frac{h}{R} \times k}$  of  $k$  2D joint points, where the third dimension  $k$  of the  $\hat{\Phi}$  vector represents the  $k$ -th joint heatmap of all person in the picture. Then input it into the 3D convolutional layer we designed for feature extraction to get  $\theta_{hm\_hp}$ . In the second branch, the RGB video sequence is directly input into the 3D convolutional layer for feature extraction, and  $\theta_{rgb}$  is obtained.

$$\theta_{fused} = \theta_{hm\_hp} \times \theta_{rgb} \quad (1)$$

After the feature extraction of the above two branches, branch1 extracted the pose heatmap  $\theta_{hm\_hp}$  with high semantic information, and branch2 extracted the RGB feature map  $\theta_{rgb}$ . We use multiplication to perform feature fusion on  $\theta_{hm\_hp}$  and  $\theta_{rgb}$ , as shown in Eq. 1. Each element of the tensor  $\theta_{hm\_hp}$  is multiplied by the corresponding element of the Tensor  $\theta_{rgb}$ . And the resulting tensor  $\theta_{fused}$  is returned.

Then, we input the linear classifier, the activation function adopts the relu function, as shown in Eq. 2. For the convergent loss function, we adopt the cross-entropy loss function, as shown in Eq. 3.

$$relu(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0 \end{cases} \quad (2)$$



**Fig. 1** The overall structure

$$H(q, p) = - \sum_x (p(x) \log q(x)) \quad (3)$$

## 4 Experiments

### 4.1 Datasets and Experimental Environment

Our experimental conditions are two 3090 GPUS, a high-performance workstation with Ubuntu 18.04. It provides hardware support for our experiment. It allows us to use CenterNet as feature extraction during the training process, and use a larger batch size to speed up the training process. Our code implementation is based on the pytorch-1.8 deep learning framework. We conducted experiments on the Ur fall detection dataset [18] and RWF-2000 [19] dataset to verify our method. For the two data sets, we use the same training settings, use a batch size of 8, a learning rate of 0.0001, and training for 30 epochs.

**Ur fall detection dataset:** contains 30 sequences of falling behavior and 40 sequences of activities of daily living (ADL). It uses two Microsoft Kinect cameras and corresponding accelerometer data to record fall events, while ADL events are recorded with only one device and accelerometer.

**RWF-2000:** Collect 2000 sequences from You Tube, which contains 1000 violent acts, 1000 ADL. All the sequences are acquired from security cameras. They have not been modified by multimedia technology, so they can be used in real-world applications. We divide the training set and the validation set according to the ratio of 8:2 for training and testing.

### 4.2 Ablation Study on UR Fall Detection Dataset and RWF-2000

**UR Fall Detection Dataset.** In order to verify the effectiveness of our transfer learning, we conducted a large number of ablation experiments. First, we only use the RGB features in branch2 for behavior recognition. The experimental results of the `rgb_only` model are shown in Table 1. The accuracy rate on the validation set of the l.

UR Fal Detection Dataset is 91.7%. In order to prove that the pose estimation information is helpful for our behavior recognition, we use CenterNet to predict the pose heatmap of the video sequence and save it locally. We directly input these

**Table 1** Results of ablation experiments on UR Fall Detection Dataset

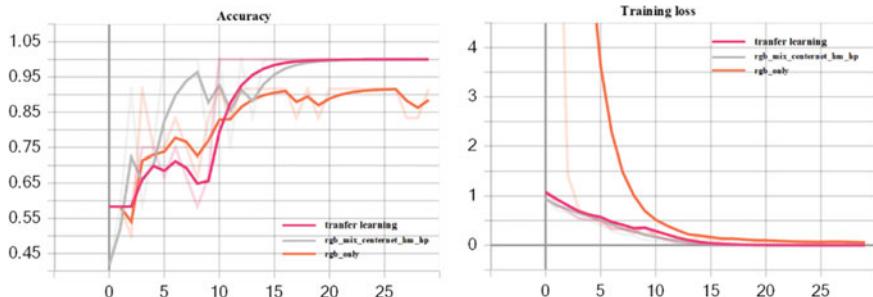
Model	Eval (%)	Test (%)
rgb_only(ours)	91.67	83.33
hm_hp_only(ours)	95.83	75.0
rgb_mix_cernet hm_hp(ours)	100.0	75.0
tranfer_learning(ours)	100.0	83.33

locally stored pose heatmaps into CNN and classifiers, and found that the accuracy of the hm\_hp\_only model in the UR Fall Detection Dataset has increased to 95.8%.

Next, we tried the rgb\_mix\_cernet hm\_hp model. After the locally stored pose heatmap and the feature map output by branch2 in Fig. 1 were multiplied and fused, they were input into CNN for prediction and found that the accuracy rate on the verification set was increased to 100% At this time, the accuracy rate on the test set is 75%. It shows that the fusion of pose heatmap and RGB can achieve better results.

We began to verify our transfer learning. As shown in Table 1, tranfer\_learning model has accuracy of 100% on the validation set, and the accuracy on the test set has increased from 75% of rgb\_mix\_cernet hm\_hp to 83.3%. It shows that we use transfer learning to get a better heatmap. At the same time, it can be concluded from the accuracy rate convergence on the validation set of the training process in left of Fig. 2 and the loss of the loss value in the training process in right of Fig. 2 that we can use migration learning to converge faster. Visualization of fall behavior detection results is shown in Fig. 3.

**RWF-2000 Dataset.** Since the UR Fall Detection Dataset is too small, we tried to perform ablation experiments on RWF-2000. As shown in Table 2, our tranfer\_learning model has a higher accuracy than rgb\_only model and rgb\_mix\_cernet hm\_hp model, reaching 83.75%. Compared with other models, our parameters are less than half of others, but the accuracy rate is higher than them.



**Fig. 2** During the training process, the accuracy rate on the validation set and the loss value change curve on the training set



**Fig. 3** Visualization of fall behavior detection results

**Table 2** The accuracy of the RWF-2000 data set on the validation set and the amount of network model parameters

model	Acc (%)	Params (M)
rgb_only(ours)	79.0	0.24843
rgb_mix_centernet_hm_hp(ours)	80.0	0.31588
tranfer_learning(ours)	83.75	20.19180
ConvLSTM [20]	77.00	47.4
C3D [21]	82.75	94.8
I3D(TwoStream) [22]	81.50	24.6

Since our network has generated a pose heatmap, in terms of visualization, we can synchronize the results of pose estimation to the visualization page at the same time, as shown in Fig. 4.

## 5 Conclusion

In this paper, We propose the transfer learning method to use pose estimation network for feature extraction, and then perform feature fusion with ordinary RGB feature maps for action detection. In addition, we conducted a large number of ablation



**Fig. 4** Violent behavior recognition and pose estimation visualization

experiments on the Ur fall detection dataset and RWF-2000 to prove that the transfer learning method we proposed is effective.

**Acknowledgements** This work was sponsored by the in part by the R & D Projects in Key Areas of Guangdong Province under Grant 2019B010109001, in part by High Resolution Earth Observation Major Project under Grant 83-Y40G33-9001-18/20, in part by Provincial Agricultural Technological Innovation and Promotion Project of Guangdong Province under Grant 2019KJ147, and in part by Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2016B030301008.

## References

1. Heng, W., Cordelia, S.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
2. Karen, S., Andrew, Z.: Two-stream convolutional networks for action recognition in videos. arXiv preprint [arXiv:1406.2199](https://arxiv.org/abs/1406.2199) (2014)
3. Limin, W., Yuanjun, X., Zhe, W., et al.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36. Springer, Cham (2016)
4. Bo, L., Yuchao, D., Xuelian, C., et al.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 601–604. IEEE (2017)
5. Du, T., Lubomir, B., Rob, F., et al.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
6. Christoph, F., Haoqi, F., Jitendra, M., et al. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)

7. Amir, S., Jun, L., Tian-Tsong, N., et al. NTU RGB+ D: a large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
8. Sijie, Y., Yuanjun, X., Dahua, L.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32(1) (2018)
9. Kaiming, H., Xiangyu, Z., Shaoqing, R., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Fisher, Y., Dequan, W., Evan, S., et al.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2403–2412 (2018)
11. Karen, S., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
12. Christian, S., Vincent, V., Sergey, I., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
13. Christian, S., Wei, L., Yangqing, J., et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
14. Shaoqing, R., Kaiming, H., Ross, G., et al. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint [arXiv:1506.01497](https://arxiv.org/abs/1506.01497) (2015)
15. Joseph, R., Ali, F.: Yolov3: an incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
16. Joseph, R.: Darknet: Open source neural networks in C. (2013)
17. Xingyi, Z., Dequan, W., Philipp, K.: Objects as points. arXiv preprint [arXiv:1904.07850](https://arxiv.org/abs/1904.07850) (2019)
18. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014)
19. Ming, C., Kunjing, C., Ming, L.: RWF-2000: an open large scale video database for violence detection. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 4183–4190 (2021)
20. Swathikiran, S., Oswald, L.: Learning to detect violent videos using convolutional long short-term memory. In: International Conference on Advanced Video and Signal Based Surveillance (AVSS 2017), pp. 1–6 (2017)
21. Du, T., Lubomir, B., Rob, F., Lorenzo, T., Manohar, P.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
22. Joao, C., Andrew, Z.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)

# Fine-Grained Visual Classification Based on Wisely Feature Map Filtering Mechanism



Haiyuan Chen , Lianglun Cheng , and Ganghan Zhang

**Abstract** Fine-Grained Visual Classification (FGVC), also named as sub-category recognition, is a research hotspot in computer vision and pattern recognition recently. The objective of FGVC is to classify objects that are in the same basic class. There are two main challenges in FGVC tasks: small interclass differences and large intraclass differences. Before the development of deep learning, there had been little significant achievement in the research of these challenges. However, most methods focus on the step of generating the feature maps but pay little attention to the further processing of feature maps. All of them are used directly by most methods after they are extracted from the original image, which lacks further processing of feature maps and may lead to irrelevant features to negatively affect network performance. In order to refine the processing of feature maps and improve the classification ability of FGVC tasks, we propose a Wisely Feature Map Filtering Network (WAMFN), which is proposed to extract the feature maps for subcategories and filter out the most distinguishable and representative feature maps to generate attention maps. Results of multi experiments show that the WAMFN outperforms the state-of-the-art methods on several fine-grained classification datasets.

**Keywords** Attention mechanism · Feature filtering · Fine-grained visual classification

## 1 Introduction

Compared to the normal image classification task, fine-grained visual classification tasks are more difficult. There are two main challenges in Fine-grained image classification: small interclass differences between subcategories and large intraclass differences. Specifically, firstly in ordinary image classification, the target objects are usually coarse-grained meta-categories. They are visually very different and easy to identify correctly. However, in fine-grained image classification, the objects to be

---

H. Chen (✉) · L. Cheng · G. Zhang  
Guangdong University of Technology, Guangzhou 510006, China  
e-mail: [2111905056@mail2.gdut.edu.cn](mailto:2111905056@mail2.gdut.edu.cn)

detected come from the same base class, which makes them very similar in appearance. At the same time, the differences between subcategories are sometimes even smaller than the same subcategory due to visual angle, which causes large intraclass differences. These two main challenges cause the difficulty for the method to learn the key features of each category as well as make it easy for other factors to influence the results, which can lead to errors in predicting the results.

Recent method of FGVC tasks can be divided into three types. The first type is strongly supervised method [1–5]. By using intensive manual annotations, local features are captured for fine-grained classification after key parts have been detected and localized. Zhang et al. [1] presented a FGVC algorithm using deep convolutional features as whole object detectors and distinguishable object detectors, and adding geometric constraints on the top of the whole and distinguishable blocks. For bird fine-grained classification, Branson et al. [2] proposed Pose Normalized CNN to perform pose alignment operations on image blocks at the part level. This work also proposed that convolutional features in different layers should be extracted for different levels of image blocks in fine-grained images. Then Wei et al. [4] proposed Mask-CNN to learn a Part-Based Segmentation Model with the help of FCN, where the real markers are the smallest external rectangles of the head and torso positions obtained through Part Annotation. We can see that strong annotations make these networks achieve good results. However, strongly supervised methods usually require additional information such as manually labeled object bounding boxes or part annotations in addition to image labels. The practical application of these methods is limited by the fact that the acquisition of labeling information is very expensive.

The second type is the semi-supervised method. These methods make use of extra data such as Internet data in training FGVC networks [6–10]. Despite increasing the data without increasing the overhead of any manual labeling, the biggest drawback of using extra data is the noisiness and uselessness of the information. Although amounts of information can be obtained from elsewhere, most of them are useless or not professional enough, and sometimes they even make the network performance degraded. It requires extra effort and cost to sift through them.

The third type is the weakly supervised method. It is a clear trend in fine-grained visual classification that making classification accuracy comparable to strongly supervised models by only using image-level annotations during training [11–26]. Recently works make some achievements. Among them, Xiao et al. [17] proposed features of two different levels, namely object-level and component-level information. The model relies entirely on its own algorithm for object and local region detection, without the requirement of datasets to provide labeling information. Besides, Chen et al. [18] proposed a network that contains two modules: destructive and constructive learning. The original image is first destructed and then reconstructed, in order to learn more distinguishing details. Moreover, WS-DAN [26] performed data augmentation through an attention mechanism. In the first stage, the input image is extracted to generate attention maps. Then, the features are augmented by maps

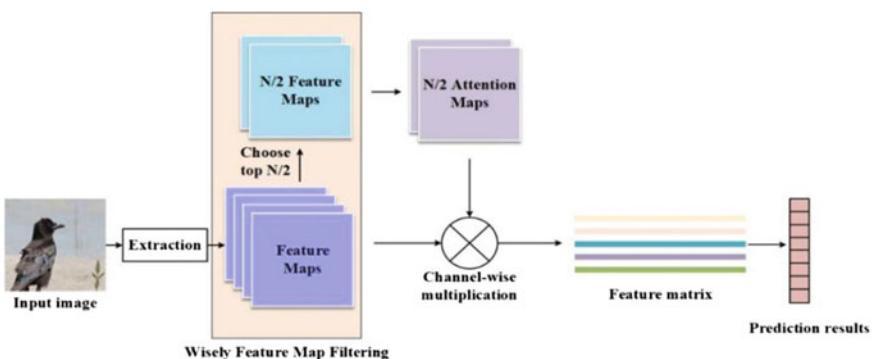
and applied to fine-tune the network in the next processing. Using a weakly supervised approach will not require significant labeling costs and is therefore becoming a popular trend in fine-grained classification.

In this work, a fast, weakly supervised and effective WAMFN is proposed to FGVC tasks. The original features extracted from the images are processed by the proposed module Wisely Feature Map Filtering to obtain the highest rated feature maps, which are used as guiding features for the images and processed to generate the attention maps. The attention maps are input into the network along with the original feature maps for processing, which enables the final network to obtain accurate prediction results.

## 2 Method

### 2.1 Overall Structure of the Network

We adopted Inception v3 as the backbone (Fig. 1). The original images are inputted into the backbone and extracted out the original feature maps by the backbone network, and then the original feature maps are filtered by the Wisely Feature Map Filtering module to filter out the best  $N/2$  feature maps with high scores, which are further processed and generate  $N/2$  attention maps. Then the original feature maps are channel-wise multiplicated with the attention maps to generate a feature matrix. Figure 2 shows the effect of the attention maps. Finally, the feature matrix is input into the fully connected layer to generate the prediction results.



**Fig. 1** Overall structure: schematic diagram of the general flow. Our network has the characteristics of simplicity, accuracy and efficiency



**Fig. 2** The effect of the attention maps: we visualize the effect of the attention maps on the dataset CUB-200-2011

## 2.2 Wisely Feature Map Filtering

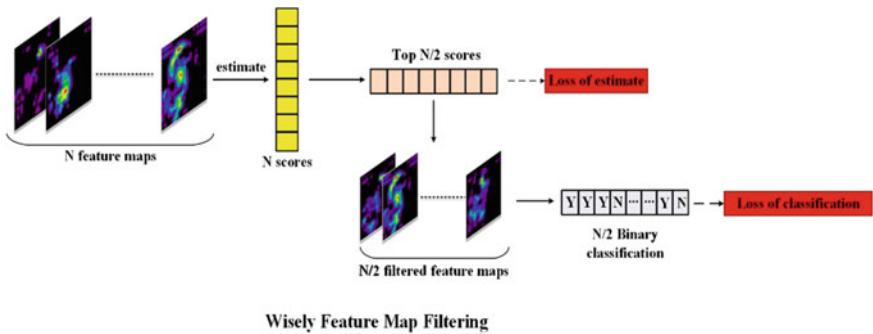
First, the original feature map  $F_o$  is inputted to Wisely Feature Map Filtering, where  $F_o$  is evaluated through convolution. Benefiting from the previous extraction of the original feature maps on the backbone network, The network is easy to calculate the score  $s_i$  of each feature map  $F_o^i$ . The original feature maps are passed through 5 convolution layers and 1 normalization layer to get the score  $S$ , as shown in Eq. (1):

$$S = [s_1, s_2, \dots, s_N] \quad (1)$$

Then all the scores are sorted and the feature maps corresponding to the top  $N/2$  scores are selected as  $F_s$ . To ensure that the filtered feature maps  $F_s$  are the high-quality maps, the network supervises and refines the process of selection through the loss of classification  $L_{cls}$  and the loss of estimate  $L_{esm}$ . We classify each filtered feature map a result  $c_i$  to determine whether it is a high-quality feature map, as shown in Eq. (2). Y (1) represents yes and N (0) represents no. Ideally, all results  $c_i$  of filtered feature maps should be Y (1). Then We use the classification results  $c_i$  to calculate the loss of classification  $L_{cls}$ , as shown in Eq. (3). At the same time, in Eq. (4), the classification results  $c_i$  was added to  $L_{esm}$  as a constraint to optimize the evaluation score for each feature map, so that high-quality attention maps would have high scores and vice versa.

$$c_i = \begin{cases} 0(N), & \text{if } \text{classify}(F_s^i) < \theta \\ 1(Y), & \text{if } \text{classify}(F_s^i) > \theta \end{cases}, \quad 1 \leq i \leq N/2 \quad (2)$$

$$L_{cls} = - \sum_{i=1}^{N/2} c'_i \log c_i, \quad c'_i = 1 \quad (3)$$



**Fig. 3** Structure of the Wisely Feature Map Filtering:  $N$  feature maps are fed into the module to select the highest quality  $N/2$  feature maps, and the selection process is automatic and intelligent by the loss function

$$L_{esm} = - \sum_{i=1}^{N/2} [(1 - c_i) \log(1 - s_i) + c_i \log s_i] \quad (4)$$

These two loss functions are optimized together during the training of the network so will eventually enable our network to have more powerful and accurate prediction results. Figure 3 is the Structure of the Wisely Feature Map Filtering.

### 3 Experiment Results

#### 3.1 Experimental Dataset and Setting

We use datasets commonly used in FGVC tasks, including CUB-200–2011 [9], Stanford Cars [10] and FGVC-Aircraft [12]. Specific information about these datasets is displayed in Table 1. CUB-200-2011 is the dataset on birds, and is also the most frequently used dataset in fine-grained visual classification and recognition. The Stanford Cars dataset is also commonly used in FGVC tasks. The FGVC-Aircraft is about the images of the aircraft in the air, which have the characteristics of long distance and different angles.

**Table 1** Detailed information on the three datasets

Datasets	Object	Class	Training images	Testing images
CUB-200-2011	Bird	200	5994	5794
Stanford cars	Car	196	8144	8041
FGVC-aircraft	Aircraft	100	6667	3333

**Table 2** The accuracy results on three datasets of comparison with other advanced methods

Methods	Accuracy (%)		
	CUB-200-2011	Stanford cars	FGVC-aircraft
RA-CNN [27]	85.4	92.5	88.4
MA-CNN [11]	86.5	92.8	89.9
DCL [18]	87.8	94.5	93.0
PA-CNN [28]	87.8	93.3	91.0
iSQRT-COV [23]	88.7	93.3	91.4
MOMN [29]	89.8	94.2	92.2
Ours	89.8	94.5	93.2

We experimented on two 2080Ti GPUs with a weight decay of 0.0001. These models had an initial momentum of 0.9, an initial learning rate was applied as 0.002, and an exponential decay of 0.8 times the original after every 2 epochs. The final settings were a training epoch of 100 and a batch size of 16.

### 3.2 Results and Analysis

The WAMFN is compared with the most recent methods on the mentioned FGVC datasets. Table 2 shows the results of accuracy. Our WAMFN has excellent performance on all three datasets, achieved the accuracy of 89.8, 94.5 and 93.2% on CUB-200-2011, Stanford Cars and FGVC-Aircraft respectively. ST-CNN and RA-CNN can locate discriminative areas to achieve FGVC task, but both rely on the bounding box provided by the dataset. compared to iSQRT-COV with a strong SVM classifier, our network achieved the highest accuracy on CUB-200-2011, Stanford Cars and FGVC-Aircraft. In a summary, the WAMFN can achieve state-of-the-art accuracy on all these FGVC datasets.

## 4 Conclusion

In this work, a new fine-grained visual classification network WAMFN is proposed. We add a Wisely Feature Maps Filtering mechanism to the generated feature maps, and extract and exploit the attention maps from the filtered feature maps, which greatly improves the quality of the classification. This module achieves better feature maps, allowing the WAMFN network to reach state-of-the-art. We hope to achieve more accurate filtering by judging the selection of attention maps according to

different base classes in the future. At the same time, we will focus on the mechanism of the network for distinguishing similar categories and expanding the network's ability to extract key differences between subcategories.

## References

1. Ning, Z., Jeff, D., Ross, G., Trevor, D.: Part-based R-CNNs for fine-grained category detection. In: Proceedings of European Conference on Computer Science, pp. 834–849 (2014)
2. Steve, B., Grant, van H., Serge, B., Pietro, P.: Bird species categorization using pose normalized deep convolutional nets. In: Proceedings of BMVC (2014)
3. Di, L., Xiaoyong, S., Cewu, L., Jiayan, J.: Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In: Proceedings of CVPR, pp. 1666–1674 (2015)
4. Xiu-Shen, W., Chen-Wei, X., Jianxin, W., Chunhua, S.: Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognit* **76**, 704–714 (2018)
5. Shaoli, H., Zhe, X., Dacheng, T., Ya, Z.: Part-stacked CNN for fine-grained visual categorization. In: Proceedings of CVPR, pp. 1173–1182 (2016)
6. Li, N., Ashok, V., Ashu, S.: Webly supervised learning meets zero-shot learning: a hybrid approach for fine-grained classification. In: Proceedings of CVPR, pp. 7171–7180 (2018)
7. Yin, C., Feng, Z., Yuanqing, L., Serge, B.: Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: Proceedings of CVPR, pp. 1153–1162 (2016)
8. Zhe, X., Shaoli, H., Ya, Z., Dacheng, T.: Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 1100–1113 (2016)
9. Zhe, X., Shaoli, H., Ya, Z., Dacheng, T.: Augmenting strong supervision using web data for fine-grained categorization. In: Proceedings of ICCV, pp. 2524–2532 (2015)
10. Jonathon, K., Benjamin, S., Andrew, H., Howard, Z., et al.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: Proceedings of ECCV, pp. 301–320 (2016)
11. Heliang, Z., Jianhong, F., Tao, M., Jiwbo, L.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of ICCV, pp. 5219–5227 (2017)
12. Chenxi, L., Linfeng, J., Jingshen, J., Weilin, Z., Huilin, X.: Weakly supervised learning of object-part attention model for fine-grained image classification. In: Proceedings of ICCT, pp. 1222–1226 (2018)
13. Ming, S., Yuchen, Y., Feng, Z., Errui, D.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of ECCV, pp. 805–821 (2018)
14. Ze, Y., Tiange, L., Dong, W., Zhiqiang, H., Jun, G., Liwei, W.: Learning to navigate for fine-grained classification. In: Proceeding of ICCV, pp. 420–435 (2018)
15. Yaming, W., Vlad, I. M., Larry, S. D.: Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proceedings of CVPR, pp. 4148–4157 (2018)
16. Tsung-Yu, L., Aruni, R., Sunhransu, M.: Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 1309–1322 (2017)
17. Tianjun, X., Yicong, X., Kuiyan, Y., Jiaxing, Z., Yuxin, P., Zheng, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of CVPR, pp. 842–850 (2015)
18. Yue, C., Yalong, B., Wei, Z., Tao, M.: Destruction and construction learning for fine-grained image recognition. In: Proceedings of CVPR, pp. 5157–5166 (2019)
19. Tsung-Yu, L., Subhransu, M.: Improved bilinear pooling with CNNs, [arXiv:1707.06772](https://arxiv.org/abs/1707.06772) (2017)
20. Yang, G., Oscar, B., Ning, Z., Trevor, D.: Compact bilinear pooling. In: Proceedings of CVPR, pp. 317–326 (2016)
21. Shu, K., Charless, F.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of CVPR, pp. 365–374 (2017)

22. Yin, C., Feng, Z., Jiang, W., Xiao, L., Yuanqing, L., Serge, B.: Kernel pooling for convolutional neural networks. In: Proceedings of CVPR, vol. 1(2), pp. 3049–3058 (2017)
23. Peihua, L., Jiangtao, X., Qilong, W., Zilin, G.: Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proceedings of CVPR, pp. 947–955 (2018)
24. Abhimanyu, D., Otkrist, G., Ramesh, R., Nikhil, N.: Maximum-Entropy Fine Grained Classification. In: Proceedings of NIPS (2018)
25. Heliang, Z., Jianlong, F., et al.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of CVPR, pp. 5012–5021 (2019)
26. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, <https://arxiv.org/abs/1901.09891>. Last accessed 23 Mar 2019
27. Jianlong, F., Heliang, Z., Tao, M.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of CVPR, pp. 4476–4484 (2017)
28. Heliang, Z., Jianlong, F., et al.: Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Trans. Image Process. 476–488 (2019)
29. Shaobo, M., Hantao, Y., Hongtao, X., Zheng-Jun, Z., Yongdong, Z.: Multi-objective matrix normalization for fine-grained visual recognition. IEEE Trans. Image Process. 4996–5009 (2020)

# Study on Prostate Image Segmentation Using Improved U-NET



Mengya Sun

**Abstract** Medical image plays a key role in the analysis and treatment of the patient's condition in today's medicine, but the clinical processing of medical image is still largely dependent on the subjective experience of doctors, and the process of manual extraction of information is time-consuming and labor-consuming. With the development of deep learning, it has become a trend to apply deep learning to medical image processing. However, due to the characteristics of medical images, many advanced segmentation algorithms fail to achieve good results in medical images. In addition, medical image data generally have the problem of insufficient or too small data sets, and there are few ways to obtain samples, mainly from some hospitals and medical institutions. Besides, the labeling work is not competent for ordinary people. So how to realize the automatic analysis and processing of medical images has always been a hot topic in the field of computer science. In this paper, we designed an experiment of prostate segmentation algorithm, built a model using a U-Net network that performs well in the field of medical image segmentation, and tried to improve its performance by combining a variety of structures, such as ASPP, ResNext, and attention module. The results show that our proposed model achieves better segmentation performance.

**Keywords** Medical image · Deep learning · Improved U-Net · ASPP · Attention(SE) · Segmentation algorithm

## 1 The Introduction

Due to the highly developed modern computer technology and the continuous development of medical imaging technology, medical imaging technology can play an important auxiliary role in a variety of research in the medical field, as well as in clinical treatment and diagnosis. Medical imaging can realize the non-invasive examination of patients in human anatomy and visibility of the development status

---

M. Sun

Department of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 610000, China

e-mail: [my18810161073@163.com](mailto:my18810161073@163.com)

of lesions, thus providing important reference and guidance for surgery, carrying out a comprehensive and in-depth analysis of lesions for tissues and organs in the human body, and carrying out a rigorous monitoring of the whole process and the treatment related to surgery. For medical analysis, it is of critical importance whether the organ tissue has a clear outline and whether there is adhesion to other tissues.

Medical image segmentation includes manual, semi-automatic and automatic segmentation [1]. Among them, the artificial segmentation method consumes a lot of time, and has high requirements on the clinician's own experience judgment, and the repeatability is very low, unable to meet a variety of clinical requirements. Semi-automatic segmentation needs human-computer interaction as the core. Generally speaking, the core methods of edge segmentation include region extraction method and other methods [2–6], which can optimize the segmentation speed to some extent, but it still needs the observers as the core for analysis and judgment. And due to the lack of self learning ability, anti-interference ability remains a low level. Therefore, it can not be applied well in clinic. Automatic segmentation method relies on the automatic extraction of the edge of the region of interest by computer, which can ensure that it is not subject to various influences of the observer on the subjective level, so as to realize more rapid data processing and have relatively ideal repeatability. Therefore, medical image segmentation with deep learning as the core is one of the core research directions for image processing at this stage [7].

The introduction of artificial intelligence in the treatment of medical influence can not only optimize the analysis and judgment of the patient's condition, promote the scientificity of treatment, but also provide scientific reference for doctors to grasp and handle the condition. At the present stage, the distribution of medical resources in China is generally uneven. If the artificial way can be introduced to improve the scientificity and rationality of medical diagnosis, the dependence on doctors' experience can be reduced, so as to avoid further aggravation of the phenomenon of medical difficulties. Therefore, the introduction of deep learning in the medical field can also make good research progress in this discipline.

With the joint efforts of a large number of researchers at home and abroad, many image segmentation methods have been widely used in the processing of medical images. Some algorithms are the optimization of some existing algorithms, or organic combination through several structures, etc. The theory of image segmentation algorithm has significant diversity. For example, the edge detection as the core [8] mainly includes parallel differential operator, the deformation model as the core and the surface fitting as the core method [9], the statistical theory as the core and the random field as the core method, etc. Generally speaking, medical image segmentation algorithms must be organically combined with a variety of existing segmentation methods at the present stage to achieve more ideal results. Whether the results are accurate or not is of critical significance for the analysis and judgment of doctors. If the results are wrong, the diagnosis of the condition by doctors will also produce errors, and even cause medical accidents. Therefore, accuracy is a crucial factor. In addition, real-time performance and stability of the algorithm need to be continuously optimized.

In recent years, medical image segmentation with deep learning as the core can be divided into two different types, namely deep convolutional neural networks (DCNNs) as the core method and full convolutional neural networks (FCNs) as the core method. FCN [10] has important performance for image segmentation, especially U-NET [11], an optimized full convolutional neural network with FCN as the core, is further transformed into the core network used for medical image segmentation. Based on IEEE retrieval, the number of papers related to it can be found to be about 200. At the same time, some of them play an important role in medical imaging and can be recognized by a large number of relevant scholars. The segmentation objects of this kind of article have high diversity, such as lung, liver and so on. The components of U-NET include an up-and-down sampling encoder and a skip connection between them. Encoding and decoding can combine local and global information organically. Up to now, U-NET has gradually produced many types of variants, all of which can play a relatively ideal role in medical image segmentation. At the same time, U-NET has been further transformed from the original 2D to 3D, thus forming new 3D U-NET [12], V-NET and other models.

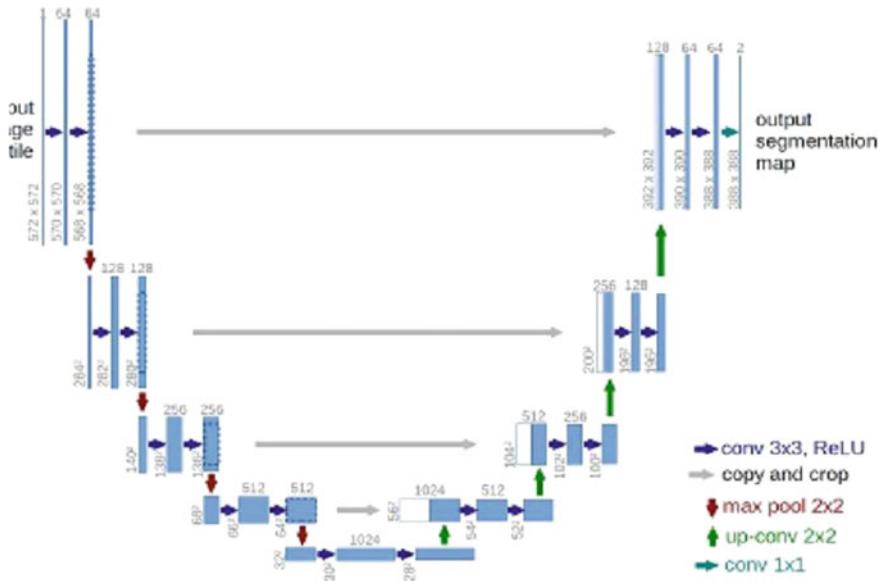
## 2 Introduction to the Foundation and Additional Structure of Improved U-NET Neural Network

### 2.1 Improve U-Net Neural Network

The core of convolutional neural network (CNN) is to learn and utilize the feature mapping of images to develop a more detailed feature mapping. It plays an obvious role in the classification problem, because the image is transformed into a vector after processing, and then the subsequent classification is expanded. However, at the same time of image segmentation, the feature image should be processed to make it into a vector, and the vector should be taken as the core to complete image reconstruction. This task is difficult because the process of turning a vector into an image is more complex than vice versa. The core concepts of U-Net are all generated in response to this problem.

Contract the path, and there will be a window of pooling equal to  $2 \times 2$  after two convolution layers, and the step length is equal to the largest pool of 2 layers. The dimension of a convolution is equal to  $3 \times 3$ , and the step length is equal to 1. All the backsides of the convolution layer needs to play the role of activation through Relu activation functions. When each one is completed under a sampling process, the number of channels increases. As for the upper sampling in the decoding path, there is a convolution layer with a size equal to  $2 \times 2$  in each round of sampling, and the activation process needs to be further completed by ReLU function (Fig. 1).

At the same time, there are also two convolutional layers, the size of the convolutional kernel is equal to  $3 \times 3$ , and the step size is equal to 1. In addition, each upsampling is organically combined with the feature graph obtained from the



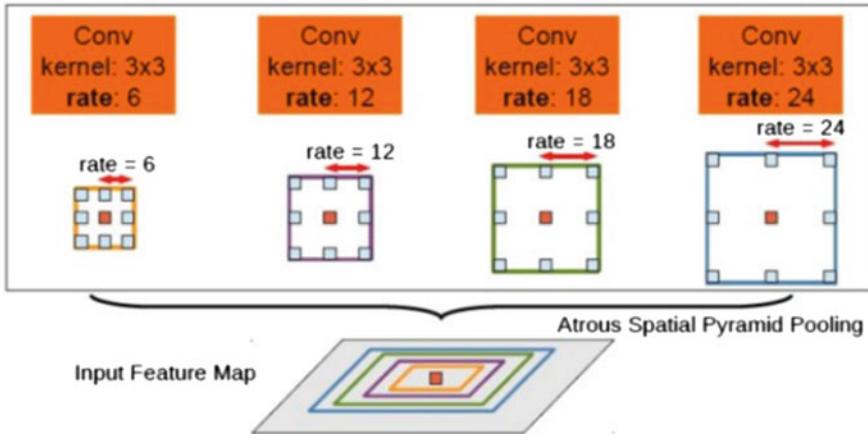
**Fig. 1** U-Net 结构

symmetric contraction path, that is, jump link. Finally, the final classification process is completed by means of a convolutional layer of size equal to  $1 \times 1$ .

This time, through the organic combination of the subsampling part contained in the U-Net network and the residual network ResNet-Layer, the corresponding transformation is carried out for U-Net.

## 2.2 ASPP

When it comes to ASPP, Dilated Convolution needs to be emphasized. Because the existing convolutional neural network has two core problems in the process of completing the segmentation task, the first is the loss of information caused by subsampling, and the second is that the convolutional neural network has the characteristic of spatial invariance. Therefore, a method is derived which can not only avoid the process of subsampling, but also effectively amplify the receptive field, namely, void convolution. In the process of convolutional pooling of the pyramid with void space, ASPP is the process of parallel sampling with the correct sampling rate and void convolution for specific inputs. In essence, it has the same property as the global information contained in the image captured by several proportions (Fig. 2).



**Fig. 2** ASPP 结构

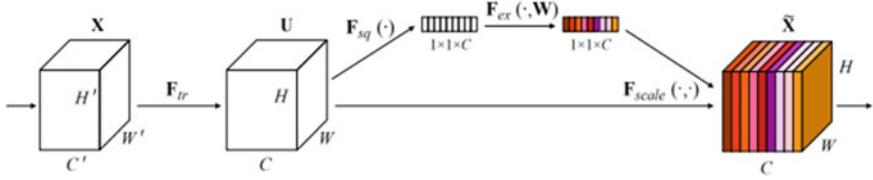
### 2.3 SE

SE block belongs to the category of the cell, and to all the specific transform structure:  $F_{tr} = X \rightarrow U$ ,  $X \in R^{H' \times W' \times C'}$ ,  $U \in R^{H \times W \times C}$ . For the purpose of simplifying the flow, in the following discussion,  $F_{tr}$  is proposed as a convolution operator. Set  $V = (v_1, v_2, \dots, v_c)$  refers to the learning set of the filter kernel, and  $v_c$  refers to the parameters of the  $c$  filter. The output of  $F_{tr}$  is expressed as  $U = (u_1, u_2, \dots, u_c)$ , and  $u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s$ .

**Extrusion: global information embedding.** In order to overcome the problem of channel dependence, a comprehensive analysis of the signal characteristics of all channels is needed. All the filters learned have the corresponding local acceptance domain. So none of the elements of the transformation output  $U$  can make use of the global information outside this region. This problem becomes more and more obvious in the lower layers of neural networks with small acceptance domains. Therefore, this study proposes to compress the global spatial information and further transfer it to the channel descriptor. Channel statistics are obtained using global average pooling. In terms of form, The statistic  $z \in R^C$  is generated by the contraction of  $U$  through the spatial dimension  $H \times W$ , and the  $C$ th element in  $Z$  is expressed as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (1)$$

**Stimulation: adaptive adjustment.** In order to make use of the aggregation information in the extrusion process, a second scrubbing was carried out to thoroughly collect the dependencies related to the channel. Therefore, the function is required to meet



**Fig. 3** SE structure

two criteria. First, it has high flexibility. Second, you need to learn a non-exclusive relationship. Therefore, the Sigmoid activation mechanism with low complexity was chosen in this study:  $s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1z))$ , where  $\delta$  refers to the Relu function.  $W_1 \in R^{\frac{C}{r} \times C}$ ,  $W_2 \in R^{C \times \frac{C}{r}}$ . Final output of the block needs to refer to scaling transformation further will be output,  $U$  finally get activation  $\tilde{X}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c$ ,  $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_C]$  (Fig. 3).

### 3 Experimental and Structural Analysis

#### 3.1 Loss Function

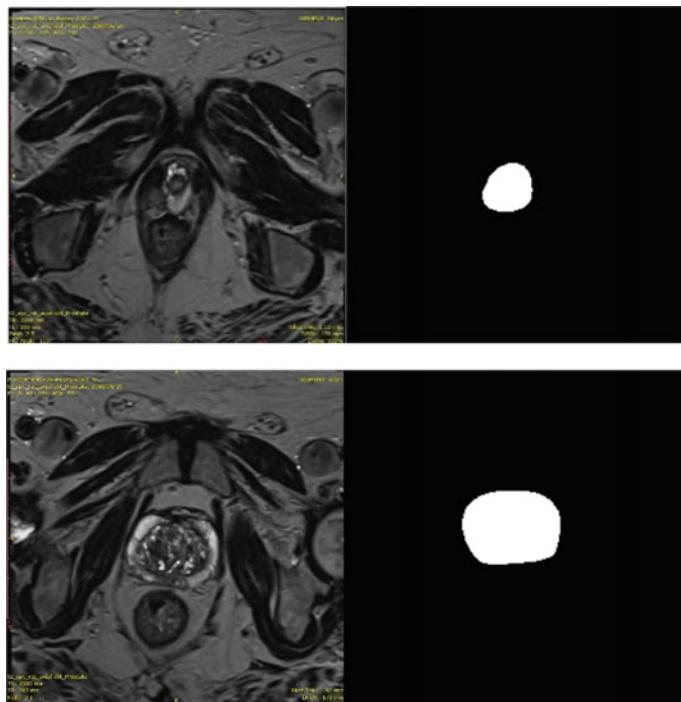
Use BCE With Logits Loss as a Loss function, which is expressed as follows:

$$l_n = -w_n [y_n \cdot \log(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))] \quad (2)$$

#### 3.2 Analysis and Comparison of Experimental Results

**The data set.** For the prostate data set used in this study, the number of training images was equal to 1549 and the number of test images was equal to 683. The processor used in the experiment was I5, the operating system was Linux, NVIDIA GeForce GTX 1650 GPU was selected, and all experimental processes were carried out based on PyTorch framework.

**Training algorithm.** The commonly used methods include stochastic gradient descent (SGD) and some adaptive training approaches, such as adaptive gradient descent (ADAGRAD), adaptive momentum estimation (ADAM), etc. The ADAM algorithm, which converges rapidly, is used in this experiment (Fig. 4 and Table 1).



**Fig. 4** Segmentation image

**Table 1** Comparison experimental results of adding different structures

	ASPP	SE	Resnet-layer	Val-Loss
U-Net				0.2020
U-Net1	✓			0.0182
U-Net2	✓	✓		0.0173
U-Net3	✓	✓	✓	0.0169

## References

1. Qian, Z.: Research and Application of Medical Image Segmentation Method. Southern Medical University, Guangdong (2014)
2. Xueming, W.: Research on Algorithms of Image Segmentation. Chengdu University of Technology (2006)
3. Hongying, H., Tian, G., Tao, L.: Overview of image segmentation methods. Comput. Knowl. Technol. **15**(5), 176–177 (2019)
4. Songjin, Y., Helei, W., Yongfen, H.: Research status and prospects of image segmentation methods. J. Nanchang Water Conservancy Hydropower Coll. **2004**(02), 15–20 (2004)
5. Lili, Z., Feng, J.: Overview of image segmentation methods. Appl. Res. Comput. **34**(07), 1921–1928 (2017)

6. Weibo, W., Zhenkuan, P.: Overview of Image segmentation methods. *World Sci. Technol. Res. Dev.* **2009**(6), 1074–1078 (2009)
7. Xiaowei, X., Qing, L., Lin, Y., et al.: Quantization of fully convolutional networks for accurate biomedical image segmentation. In: Proceedings of the 36th International Conference on Computer Vision and Pattern Recognition, pp. 8300–8308. IEEE Press, NJ (2018)
8. Djemel, Z., Salvatore, T.: Edge detection techniques—an overview. *Int. J. Pattern Recogn. Image Anal.* **8**(4), 537–559 (1998)
9. Alvarez, L., Morel, L.: Image selective smoothing and edge detection by nonlinear diffusion. II. *Siam J. Numer. Anal.* **29**(3), 845–866 (1992)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2015)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, Berlin (2015)
12. Iek, Z., Abdulkadir, A., Lienkamp, S.S., et al.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Springer, Cham (2016)

# Music Auto-tagging Based on Attention Mechanism and Multi-label Classification



Chen Ju , Lixin Han , and Guozheng Peng

**Abstract** Music tag is used to describe music, which is the key of music information retrieval and music recommendation system. Aiming at the problems of heavy manual workload and difficult audio feature extraction existing in traditional methods, this paper proposes an audio automatic annotation method based on time–frequency domain analysis, and conducts experiments on MagnaTagATune dataset with AUC-ROC as the evaluation index. Firstly, we use the modal translation transform the music audio into Mel-spectrum. The convolutional neural network is used to learn the time–frequency domain characteristics of the audio signal. Simultaneously, attention mechanism is introduced to broaden the receptive field of the feature graph and enhance the learning of the global spatial dependence. In addition, the loss function is optimized based on the image segmentation problem to improve the performance. The experimental results show that our proposed model is effective and robust in music auto-tagging, and it can achieve the ROC-AUC score of 91.87%.

**Keywords** Music auto-tagging · Modal translation · CNN · Attention mechanism · Classification loss function

## 1 Introduction

Music auto-tagging [1] is a classification task with vast labels predicted from audio signals, such as genre, instrument, emotion, etc. From the perspective of listeners, labels are high-level descriptive words to express music characteristics, so music automatic tagging task would be the crucial part in music retrieval, music visualization, music recommendation and so on [2].

Over the last few years, much attention has been put on deep learning approaches [3, 4]. The establishment of auto-tagging model can be divided into three parts, preprocessing, extracting feature and classifying. The extracted features which are time-consuming and require very professional prior knowledge are heavily relied

---

C. Ju · L. Han · G. Peng ()  
HoHai University, Nanjing 211106, China  
e-mail: [jerrypeng@hhu.edu.cn](mailto:jerrypeng@hhu.edu.cn)

on in traditional machine learning. As a rule, the result is closely related to the quality of features. But with the latest breakthrough in deep neural networks, the paradigm has been shifted to learning representations with the original waveform or time-frequency representation as the input.

Convolutional neural network (CNN), one of the most popular approaches, is capable of learning invariant features of spatial layout automatically from bottom to top [5] and have achieved great performances in music auto-tagging tasks. Hamel [6] carried out deep belief networks for feature extraction in an unsupervised way. Pons [7] discussed the effects of CNN using waveform and spectrogram as inputs in large-scale data scenarios. Lee et al. [3, 8] proposed two new structures, namely the global model with both multi features and Sample-CNN, which obtained the AUC-ROC score of 90.17% and 90.55% on MagnaTagATune (MTAT) dataset [9] respectively, but the networks they used are very deep.

Although CNN-based models can accomplish music automatic labeling task very well, there are still some improvements to be made. One approach is to enhance the feature extraction capability. On this basis, we propose a novel model with attention mechanism applied to generate potential spatial focus locations. Except for this idea, we introduce label correlations for optimization since there may be abundant relationships among rich labels of the same instance. In this paper, we conduct experiments on MTAT datasets, and the result shows that, compared with other studies, our proposed model is remarkable in music auto-tagging task.

## 2 Related Work

### 2.1 Modal Translation

Modal translation is a core technology in multi-modal machine learning. Converting data from one schema to another, generating different schemas of the same entity [10], which can better optimize the target, which has important applications in speech recognition and synthesis, visual scene description, cross-modal retrieval and other fields.

In this paper, we use modal translation to transform the audio mode of music into the image mode of spectrogram [11], where the abscissa axis is time, the vertical axis means frequency, and the shade of color indicates the energy of the sound. At present, spectrogram has become an important method for audio analysis because of its simpler and more compact expression.

## 2.2 Attention Mechanism and Self-attention

Now, CNN has become the preferred method to encode image signals because its learning process mimics the human visual system. CNN mainly relies on convolution operation, using local receptive field to integrate spatial information and channel information to extract features. For further improvement, attention mechanism [12] can be introduced to selectively strengthen useful features and suppress features containing useless information by learning from a global perspective.

Considering that the dependencies between channels are not taken into account during the convolution process, we add Squeeze-and-Excitation (SE) block [13] to extend the model. As Fig. 1 shows, SE block is mainly made up of two operations: squeeze and excitation.

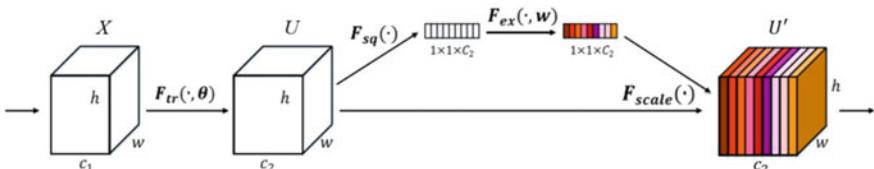
First, the global spatial information is squeezed into channel descriptors through a global average pooling layer and statistics for each channel are generated. In detail, a statistic  $z \in \mathbb{R}^{c_2}$  is generated by compressing  $U$  through  $h \times w$  spatial dimension, hence, the  $c$ -th element of  $z$  can be computed as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{h \times w} \sum_i^h \sum_j^w u_c(i, j) \quad (1)$$

where  $U = [u_1, u_2, \dots, u_{c_2}]$  and  $u_c \in \mathbb{R}^{h \times w}$ . Then the weight of each channel is obtained by sigmoid function [14], according to which the resolution of the feature is enhanced. Given  $z$ , the output  $s$  can be expressed as:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \quad (2)$$

where  $\sigma$  refers to sigmoid function while  $\delta$  is ReLU function [15],  $W_1 \in \mathbb{R}^{\frac{c_2}{r} \times c_2}$  and  $W_2 \in \mathbb{R}^{c_2 \times \frac{c_2}{r}}$  are learnable, and  $F_{scale}(\cdot)$  is the product between the scalar  $s_c$  and the feature map  $u_c$ . Thus, it can be seen that SE block intrinsically introduces dynamics conditioned on the input that can be regarded as an automatic attention focusing on channels and is not limited to the local receive region of the transform filter response. On the other hand, it greatly improves the capability with only a small increase in computing consumption.



**Fig. 1** SE block

### 2.3 Multi-label Classification

Audio automatic annotation is a multi-label classification where every instance is associated with a set of tags and the value of each label is either 0 or 1, depending on whether the audio has been assigned the tag. Actually, in the real world one instance may belong to a mass of labels simultaneously so that multi-label classification is more common and important, which attracts extensive attention in image classification [16] and text categorization [17]. In order to address the problem, researchers have put forward various ways to classify multi labels including first-order, second-order and high-order strategies.

In the neural network, the probability of the sample  $x_i$  being tagged with the  $j$ -th label is modeled as Bernoulli probability distribution by performing sigmoid function on the output of the neural network, and the calculation is:

$$P(y_j|x_i) = \frac{1}{1 + \exp(-z_j)} \quad (3)$$

where  $z_j$  represents the  $j$ -th eigenvalue of the output  $Z$  of the network, that is,  $Z = [z_1, z_2, \dots, z_n]$ , while  $n$  means the number of tags. Thus, the probability distributions of each category are independent from each other. Then, binary cross entropy (BCE) [18] is used to calculate the error between predicted value and true value, as shown below:

$$L_{BCE} = \frac{1}{m} \sum_m \left( \left( -\frac{1}{n} \right) * \sum_n (y_i \times \ln(p_i) + (1 - y_i) \times \ln(1 - p_i)) \right) \quad (4)$$

where  $m$  represents the batch size,  $n$  is the total number of tags,  $y_i \in \{0, 1\}$  is the value of  $i$ -th label, and  $p_i \in [0, 1]$  is the probability predicted as  $i$ -th label. However, an inevitable drawback in this method is that the correlations among labels are not taken into account. As a matter of fact, there may be abundant relationships among vast labels of the same instance. Thus, it is important to consider it for further improvement.

In this paper, we ameliorate the problem by introducing the correlation among tags, which can be expressed in terms of similarity calculations between sets. Given  $X, Y$  as two sets, the similarity can be calculated as follows:

$$S(X, Y) = \frac{2 * |X \cap Y| + 1}{|X| + |Y| + 1} \quad (5)$$

where  $|X \cap Y|$  is implemented by taking the dot product of the matrix elements and then adding them up, while  $|X|$  is obtained by summing the square of its elements, the same to  $Y$ . Otherwise, we add 1 to keep the denominator from zeroing out. During this operation, the tags are associated. The correlation loss can be defined as:

$$L_{Cor} = \frac{1}{m} \sum_m (1 - S(P_i, Y_i)) \quad (6)$$

where  $m$  is batch size,  $P_i$  is the predicted label set of sample  $x_i$  while  $Y_i$  is the true label set. Finally, the loss function can be expressed as:

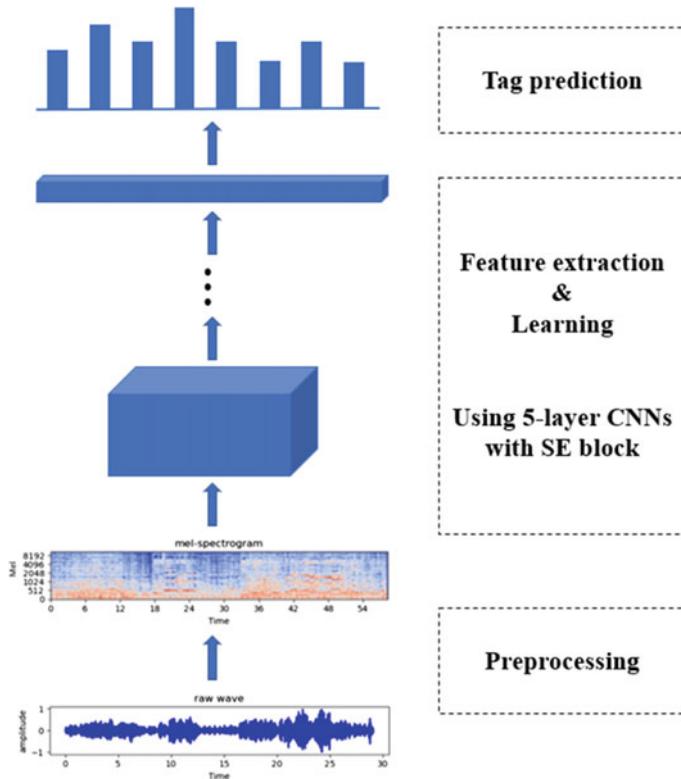
$$L = \alpha L_{BCE} + (1 - \alpha) L_{Cor} \quad (7)$$

where  $\alpha$  is a hyperparameter that measures the importance of the two losses. Therefore, we can not only close the distance between predicted label set and true label set from the micro level but also investigate it from the global level based on the correlation among tags.

### 3 Model

As figured out in Fig. 2, our proposed model mainly consists of three stages.

Firstly, raw music data are transformed into mel-spectrogram. Mel-spectrogram is a comprehensive representation which can show the dynamic change of sound frequency and energy over time. In the second part, the network composed of 5 convolution layers is employed to learn the time–frequency domain characteristics at a granular level and the details are shown in Table 1. Each convolution layer uses ReLU activation responsible for enhancing the non-linearity of the network and SE block dynamically focusing on the important parts on the channel dimension. To reduce the computational load while retaining the feature space information, we use the max-pooling layers, reducing overfitting and improving the robustness and fault tolerance of the model. After the last convolution layer, we flat the feature maps into a vector and use batch normalization to reduce internal covariate offsets. Furthermore, dropout layer with probability set as 0.6 is added to avoid overfitting. Finally, the sigmoid-activated output is used as the confidence level for each text label that the model predicts for the music sample. At the third stage, calculate the error between the prediction and the true label set according to the loss function  $L$  given in Sect. 2.3 and the error is propagated back to update the network parameters, making the predicted value constantly approaching the real value. By the way, the model is built with pytorch framework in a GPU environment and batch size is 20.



**Fig. 2** The structure of our proposed model

## 4 Experiments

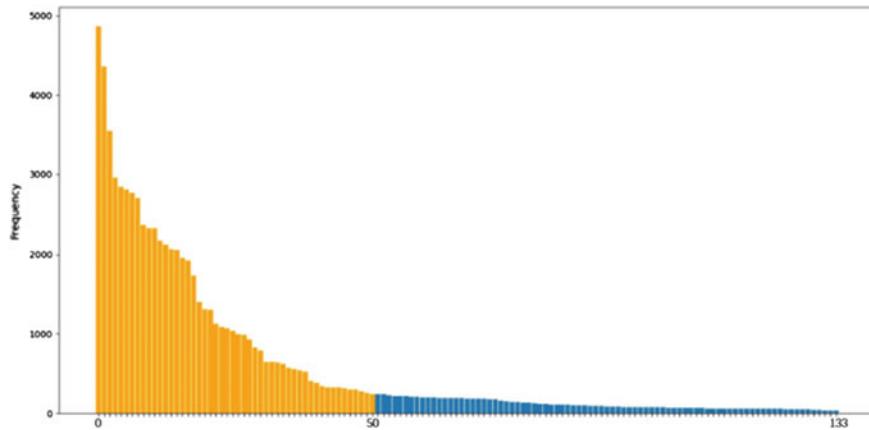
### 4.1 Dataset and Preprocessing

We evaluate our proposed model on MTAT dataset, which is one of the most popular publicly available datasets for music auto-tagging analysis. It contains 25,863 music clips in .mp3 format, each with 188 tags covering genres (e.g., country, metal), instruments (e.g., violin, guitar), scene (e.g., choral, vocal), etc.

Since some music tags are synonyms that have different names but the exact meaning is the same, we firstly merge the synonym tags and reduce the number to 133. Figure 3 shows the frequency of each label after processing, and we can see that the tags are not evenly distributed. It is significantly hard to train any deep learning model to classify tags with a small sample size. Hence, same as other existing research work, we use the top 50 most frequently occurring tags for our experiments. The audio data is cleaned and resampled to 11,025 Hz by Librosa [13]. There are 21,350 fragments, the duration of each is 29.124 s. Each MP3 signal is converted into

**Table 1** Details of the CNN layers

Layer	Setting	Output shape
Conv2d_1	Filter size: (3, 7). Number of filters: 50	(128, 628, 50)
SE_1	Channel of feature map: 50 Reduction ratio: r	(128, 628, 50)
MP_1	Filter size: (2, 4)	(64, 157, 50)
Conv2d_2	Filter size: (3, 5). Number of filters: 100	(64, 157, 100)
SE_2	Channel of feature map: 100 Reduction ratio: r	(64, 157, 100)
MP_2	Filter size: (2, 4)	(32, 39, 100)
Conv2d_3	Filter size: (3, 3). Number of filters: 70	(32, 39, 70)
SE_3	Channel of feature map: 70 Reduction ratio: r	(32, 39, 70)
MP_3	Filter size: (2, 2)	(16, 19, 70)
Conv2d_4	Filter size: (3, 3). Number of filters: 70	(16, 19, 70)
SE_4	Channel of feature map: 70 Reduction ratio: r	(16, 19, 70)
Conv2d_5	Filter size: (3, 3). Number of filters: 70	(16, 19, 70)
SE_5	Channel of feature map: 70 Reduction ratio: r	(16, 19, 70)
MP_5	Filter size: (2, 2)	(8, 9, 70)

**Fig. 3** Frequency of each label (yellow is the top 50 tags and blue is the rest)

**Table 2** Results of different location of the SE block

	Reduction ratio	AUC-ROC score	Time (h)
After each convolution	2	0.9096	1.72
	4	0.9095	1.73
	8	0.9115	1.75
	16	0.9121	<b>1.66</b>
	32	<b>0.9158</b>	1.69
After the last convolution	2	0.911	1.61
	4	0.911	1.59
	8	0.9106	<b>1.57</b>
	16	<b>0.9133</b>	1.61
	32	0.9106	1.58

mel-spectrogram as the input of the network through short-time Fourier transform and mel-filter bank, with the size of (128, 628, 1), which respectively represents frequency, time and channel. Besides, 15,265 clips are chosen randomly for training, 1520 for validating and 4565 for test, and Adam [19] optimization with learning rate of 0.0001 is used in training process.

## 4.2 Results and Analysis

The architecture is explored with the AUC-ROC score as the evaluation indicator. Since the location of the SE block makes difference in the performance, we conduct the experiment by placing it after each convolution and after the last convolution, and the results are shown in Table 2. In a word, with the number of SE blocks used in the network, the performance rises, but so does time consumption.

Allowing a certain amount of time consumption, we add SE block after each convolution layer for further research. When the self-attention is paid to the channels of the feature map obtained from the convolution operation, the reduction ratio is used to compress the channel dimension so that the ReLU and sigmoid function can be applied to obtain the complex correlation between channels, which is usually valued as 2, 4, 8, 16, 32. The influence of its value on the experimental results can be observed in Table 3, where Time represents the total of training and testing time. When reduction ratio is 16, the performance is improved by 0.83% with the lowest time consumption, while, when the value is 32, the proposed model-1 (CNN with SE block, using BCE loss) can achieve the AUC-ROC score of 91.58%, with a 1.4% improvement compared to model-0 (the network without the attention mechanism). It's worth noting that the time is only increased by 7.6%. According to the loss function  $L$  given in Sect. 2.3, correlation among tags has been considered to further improve the classifier's performance. As we can see, the highest AUC-ROC score

**Table 3** Results of different reduction ratio

	Reduction ratio	AUC-ROC score	Time (h)
Model-0	—	0.9018	1.57
Model-1	2	0.9096	1.72
	4	0.9095	1.73
	8	0.9115	1.75
	16	0.9121	<b>1.66</b>
	32	<b>0.9158</b>	1.69
Model-2	2	0.9125	1.87
	4	0.9095	1.87
	8	0.9137	<b>1.86</b>
	16	0.9135	1.89
	32	<b>0.9187</b>	1.88

**Table 4** Comparison of proposed architecture to previous methods on MATA dataset

Input type	Model name	AUC-ROC score
Raw waveform	Sample CNN [8]	0.9055
Raw waveform	DCNN [20]	<b>0.9276</b>
Spectrogram	Global model with both multi-features [3]	0.9021
Spectrogram	Capsule network [21]	0.9067
Spectrogram	Our model	0.9187

achieved is 91.87%. Although, it is not a big boost, the performance of model-2 is superior to model-1 as a whole. Compared with base model (model-0), there is a 1.7% point in increase.

The proposed model is also compared with other researches conducting experiments on the MATA dataset, and the results are summarized in Table 4, which suggests that our proposed model gets the higher AUC-ROC score than most of them. In general, our proposed model with self-attention focused on the channel dimension of the feature map after each convolution layer works better and the introduction of associations among can help classify music labels more effectively.

## 5 Conclusion

In this paper, we proposed a CNN-based model with which takes spectrogram as input for music auto-tagging. By comparison, the performance of our model is more excellent on the MTAT dataset, which implies the combination of convolutional layer and SE block can enhance the capability of feature extraction. And, the introduction

of associations among tags can help classify multi-label more effectively to a certain extent. In the future, taking different forms of music as input will be investigated. On the other hand, we will use CRNN to learn the spatiotemporal relationship in a spectrogram.

## References

- Shaleen, B., Vadivel, S., Arul, J.J.: Efficient music auto-tagging with convolutional neural networks. *J. Comput. Sci.* **15**(8), 1203–1208 (2019)
- Jiao, P., Yang, Y.: Music annotation and retrieval using unlabeled exemplars: correlation and sparse codes. *IEEE Signal Process. Lett.* **22**(10), 1771–1775 (2015)
- Lee, J., Nam, J.: Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE Signal Process. Lett.* **24**(8), 1208–1212 (2017)
- Jialien, H., ChienChang, H.: Designing a graph-based framework to support a multi-modal approach for music information retrieval. *Multimedia Tools Appl.* **74**(15), 5401–5427 (2015)
- Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: 2014 European Conference on Computer Vision, pp. 818–833. Springer, Berlin (2014)
- Hamel, P., Eck, D.: Learning features from music audio with deep belief networks. In: 11th International Conference on Music Information Retrieval (ISMIR), Utrecht, The Netherlands, pp. 339–344 (2010)
- Pons, J., Nieto, O., Prockup, M., Schmidt, E.M., Ehmann, A.F., Serra, X.: End-to-end learning for music audio tagging at scale. In: 19th International Conference on Music Information Retrieval (ISMIR), Paris, France, pp. 637–644 (2018)
- Lee, J., Park, J., Kim, K.L., Nam, J.: Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. In: ICASSP, pp. 366–370. IEEE, Canada (2017)
- Edith, L., Kris, W., Michael, M., Mert, B., Stephen, D.: Evaluation of algorithms using games: the case of music annotation. In: 10th International Conference on Music Information Retrieval (ISMIR), Kobe, Japan, pp. 387–392 (2009)
- Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 423–443 (2019)
- Qing, C., Qian, G., Ming, Z.: Analysis of vocalicity on Spectrogram. *Microcomput. Inf.* **26**(21), 6–8 (2010)
- Zhen, C., Maoyong, C., Peng, J., Fengying, M.: Research on crop disease classification algorithm based on mixed attention mechanism. *J. Phys. Conf. Ser.* **1961**(1), 1–7 (2021)
- Jie, H., Shen, L., Albanie, S., Gang, S., Enhua, W.: Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 1–13 (2020)
- Langer, S.: Approximating smooth functions by deep neural networks with sigmoid activation function. *J. Multivar. Anal.* **182**(5), 1–21 (2021)
- Mohit, A., Suneet, G., Biswas, K.: A new Conv2D model with modified ReLU activation function for identification of disease type and severity in cucumber plant. *Sustain. Comput. Inf. Syst.* **30**(5), 1–17 (2021)
- Qiang, L., Maoying, Q., Wei, B., Dacheng, T.: Conditional graphical lasso for multi-label image classification. In: 2016 CVPR, Las Vegas, USA, pp. 2977–2986 (2016)
- Hossain, M.R., Hoque, M.M., Siddique, N., Sarker Iqbal, H.: Bengali text document categorization based on very deep convolution neural network. *Expert Syst. Appl.* **184**(5), 1–23 (2021)
- Tarekegn, A., Giacobini, M., Michalak, K.: A review of methods for imbalanced multi-label classification. *Pattern Recogn.* **118**(4), 1–10 (2021)
- Imran, J., Amelia, R.I., Syed, Q.N.: Adam optimization algorithm for wide and deep neural network. *Knowl. Eng. Data Sci.* **2**(1), 41–46 (2019)

20. Yongbin, Y., Minhui, Q., Yifan, T., Quanxin, D., et al.: A sample-level DCNN for music auto-tagging. *Multimedia Tools Appl.* **80**, 11459–11469 (2021)
21. Yongbin, Y., Yifan, T., Minhui, Q., Feng, M., Quanxin, D.: Music auto-tagging with capsule network. In: 6th International Conference of Pioneering Computer Scientists, Engineers and Educators (ICPCSEE), pp. 292–298. CCIS, Taiyuan, China (2020)

# Prediction of Apoplexy Syndrome Based on Graph Neural Network



Shuoyan Zhang , Zhuangzhi Yan , Jiehui Jiang , and Tianyu Gu

**Abstract** Apoplexy is a serious disease with high mortality and disability rate. Accurate identification of the syndrome of apoplexy is a prerequisite for traditional Chinese medicine treatment. Syndrome differentiation is a comprehensive diagnosis of the patient's symptoms by traditional Chinese medicine physicians. However, the relationship between syndromes and symptoms is complex, which makes it difficult to differentiate syndromes. For example, the same symptom occurs simultaneously in different syndromes, and a combination of several symptoms identifies a syndrome. Fortunately, graph neural networks provide an effective way to deal with such complex relationships. In this study, we use the graph neural network models to predict apoplexy patient's syndrome. Compared with the classical machine learning model, the graph neural network model achieves better experimental results. In addition, to our knowledge, this study is the first time that graph neural network has been applied to syndrome differentiation.

**Keywords** Traditional Chinese medicine · Syndrome differentiation · Graph neural network

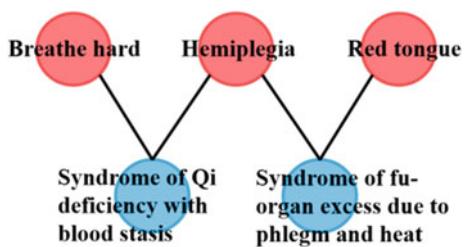
## 1 Introduction

Apoplexy is a serious threat to human health and has a high mortality and disability rate. Traditional Chinese Medicine (TCM) has the advantages of efficiency and less palindromia in the treatment of apoplexy sequelae. The basic principle of TCM is the treatment based on syndrome differentiation, which is a comprehensive examination of the patient's symptoms through the four diagnosis, and to come up with a syndrome describing the pathogenesis and location of the disease. There are many syndromes of apoplexy in TCM, such as syndrome of Qi deficiency with blood stasis, syndrome of fu-organ excess due to phlegm and heat, etc. However, syndromes and symptoms cross each other, and the same symptom may appear simultaneously in different

---

S. Zhang · Z. Yan · J. Jiang · T. Gu  
Shanghai University, Shanghai 200444, China  
e-mail: [zhangshuoyan@shu.edu.cn](mailto:zhangshuoyan@shu.edu.cn)

**Fig. 1** Complex relationship between syndromes and symptoms



syndromes, forming a TCM apoplexy syndrome-symptom graph network, as shown in Fig. 1. The nodes in red are symptoms, the nodes in blue are syndromes. Due to the complex relationship between multiple syndromes and symptoms, it is more difficult to differentiate syndromes.

Recently, the emergence of graph neural network provides a new idea for processing non-Euclidean space data such as TCM apoplexy syndrome-symptom graph network. The syndrome-symptom graph network is used as the input by the neural network to learn the representation vector of the nodes in the graph through the message transmission between neighboring nodes, and the embedding among neighboring nodes has a similar vector representation. Therefore, the node of syndrome in the syndrome-symptom graph network has a similar vector representation to its corresponding symptom node, and such similar relationship can represent prior knowledge. For example, in Fig. 1, ‘Breathe hard’ and ‘Syndrome of Qi deficiency with blood stasis’ have similar embedding representation, while ‘Red tongue’ and ‘Syndrome of fu-organ excess due to phlegm and heat’ have similar embedding representation. In this way, the embedding representation of ‘Breathe hard’ as the feature vector potentially points to syndrome of Qi deficiency with blood stasis, rather than to syndrome of fu-organ excess due to phlegm and heat.

With the development of artificial intelligence, machine learning as the core part of artificial intelligence, has been more and more applied in TCM aided diagnosis [1]. Xie et al. [2] used artificial neural network, K-nearest neighbor, support vector machine, decision tree, random forest, and Adaboost algorithm to classify the four syndromes of rheumatoid arthritis. Yan et al. [3] used support vector machine, BP neural network and extreme learning machine to classify the syndromes of 670 medical records. Pang et al. [4] used naive bayes, support vector machines, random forests, multilayer perceptron to classify the seven syndrome of AIDS.

In this research, the description of apoplexy syndromes in the book are constructed into a graph network as the prior knowledge, and the graph neural network is used to predict the patient’s syndrome.

## 2 Methods

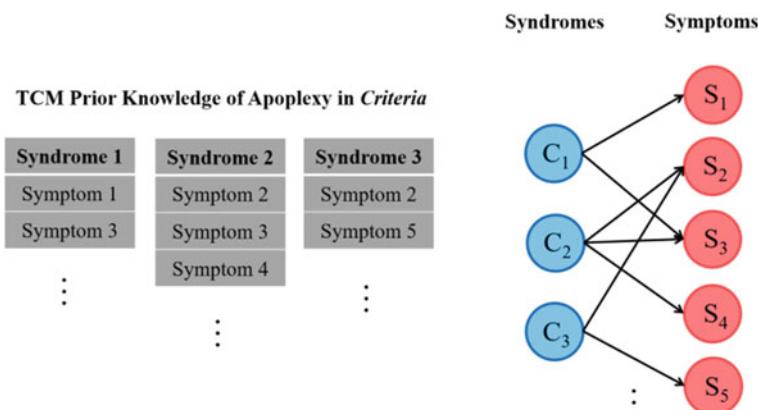
### 2.1 Basic Notation

We assume that the TCM apoplexy syndrome-symptom graph network  $\mathcal{G}_1$  expresses the prior knowledge of syndrome differentiation. Graph network  $\mathcal{G}_1$  is represented as a directed graph whose nodes form a set  $\mathcal{V}_1 = \mathcal{V}^C + \mathcal{V}^S$ , set  $\mathcal{V}^C$  consists of TCM syndromes of apoplexy, set  $\mathcal{V}^S$  consists of symptoms of apoplexy. Set  $\mathcal{E}_1$  denotes the edges between the  $\mathcal{V}^C$  and  $\mathcal{V}^S$ .

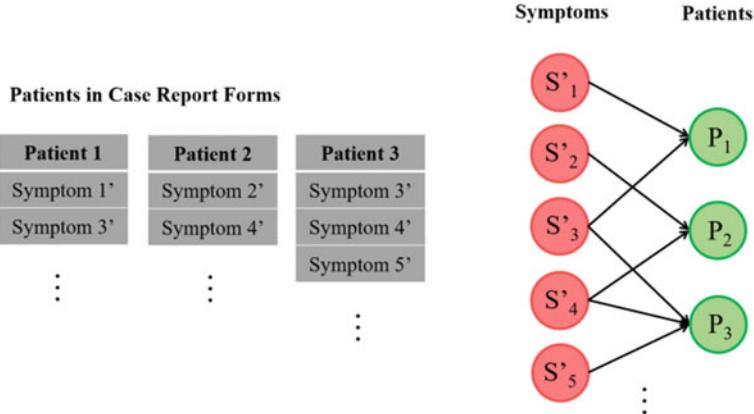
We denote the apoplexy symptom-patient graph network  $\mathcal{G}_2$  expresses the patients and their symptoms in case report forms. Graph network  $\mathcal{G}_2$  is also represented as a directed graph whose nodes form a set  $\mathcal{V}_2 = \mathcal{V}^P + \mathcal{V}^{S'}$ ,  $\mathcal{V}^P$  consists of the patients from case report forms. Due to the clinical diversity, there are some differences between the symptoms of patients in case report forms and the symptoms description of prior knowledge, so we use  $\mathcal{V}^{S'}$  to represent the symptoms correspond to patients. And set  $\mathcal{E}_2$  denotes the edges between the  $\mathcal{V}^P$  and  $\mathcal{V}^{S'}$ .

### 2.2 Building the Graph Network

In this section, we explain how to build the TCM apoplexy syndrome-symptom graph network  $\mathcal{G}_1$  and the symptom-patient graph network  $\mathcal{G}_2$  [5]. We construct the TCM apoplexy syndrome-symptom graph network based on the description of apoplexy syndromes in the *Criteria of diagnosis and therapeutic effect of internal diseases and syndromes in traditional Chinese medicine* (ZY/T 001.1-94), also known as *Criteria*. As shown in Fig. 2, on the left, syndromes and symptoms come from apoplexy



**Fig. 2** Building the TCM apoplexy syndrome-symptom graph network



**Fig. 3** Building the TCM apoplexy symptom-patient graph network

description in *Criteria*. Syndrome 1 is associated with Symptom 1, Symptom 3, on the right,  $C_1$  represents Syndrome 1,  $S_1$  and  $S_3$  represents Symptom 1 and Symptom 3 respectively. Because the *Criteria* describes the symptoms according to the syndrome, so we point the syndrome nodes to the symptom nodes. We connect  $C_1$  with  $S_1$  and  $S_3$  by directed edges. In the above way, the apoplexy syndrome-symptom graph network is constructed.

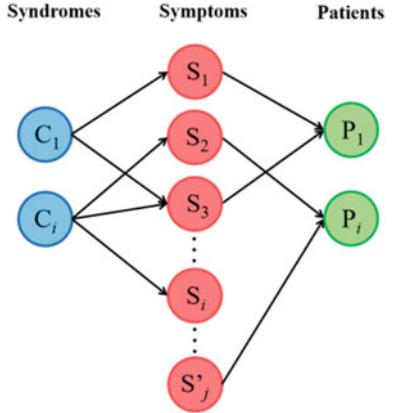
The TCM apoplexy symptom-patient graph network  $\mathcal{G}_2$  is constructed through the case report forms. As shown in Fig. 3, on the left, Patient 1 has Symptom 1', Symptom 3', on the right,  $P_1$  represents Patient 1,  $S'_1$  and  $S'_3$  represent Symptom 1' and Symptom 3' respectively. Because doctors judge patients' syndromes based on their symptoms, so we point the symptom nodes to the patient nodes. We connect  $P_1$  with  $S'_1$  and  $S'_3$  by directed edges. In the above way, the apoplexy symptom-patient graph network is constructed.

After  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are constructed, the node set  $\mathcal{V}^S$  and node set  $\mathcal{V}^{S'}$  have many same symptom nodes, so we merge these same symptom nodes. Meanwhile,  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are merged into one graph  $\mathcal{G}$ .  $\mathcal{G}$  is shown in Fig. 4. If symptom node from  $\mathcal{G}_1$  and symptom node from  $\mathcal{G}_2$  are the same, the latter symptom node would be replaced by the former symptom node. Besides, in Fig. 4, node  $S_i$  represents the symptom only in  $\mathcal{G}_1$ , node  $S'_j$  represents the symptom only in  $\mathcal{G}_2$ . In the following work, our predicting task will be calculated on this one graph  $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$ ,  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  and  $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2$ ,  $\mathcal{V} = \mathcal{V}^C + \mathcal{V}^P + \mathcal{V}^S + \mathcal{V}^{S'}$ .

### 2.3 Graph Neural Network Model

The predicting the apoplexy syndrome model takes the graph network  $\mathcal{G}$  as its input, and generates the embedding vectors of nodes through the aggregate function which

**Fig. 4** Merging the  $\mathcal{G}_1$  with  $\mathcal{G}_2$



calculate the interaction of neighbor nodes. Then, given a patient, we can get an embedding vector of patient by fusing the embedding vector of symptoms in his or her case report form. Finally, we use this embedding estimate the patient with the syndrome.

We select the popular graph neural network models such as GraphSAGE [6], GCN [7] and RGCN [8] to predict the syndromes. Given a graph  $\mathcal{G}$ , we denote a syndrome, symptom or patient as  $u$  or  $v$ ,  $u \in \mathcal{V}$  and  $v \in \mathcal{V}$ .  $k$  represents the  $k$ -th information aggregation on the graph.  $\mathbf{h}_v^k$  represents the embedding vector of the node  $v$ .

GraphSAGE model calculates as follows:

$$\mathbf{h}_{\mathcal{N}(v)}^k = \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\}) \quad (1)$$

$$\mathbf{h}_v^k = \sigma(\mathbf{W}^k [\mathbf{h}_v^{k-1}; \mathbf{h}_{\mathcal{N}(v)}^k]) \quad (2)$$

where  $u$  represents the neighbor node of the  $v$ ,  $\mathcal{N}(v)$  denotes the neighbor set of the  $v$ .  $\text{AGGREGATE}_k$  represents the aggregate function, such as Recurrent Neural Network, mean pooling, max pooling, we adopt mean pooling for information aggregation. We use  $[\cdot; \cdot]$  to represent the concatenation of two vectors. The parameter matrix is represented by  $\mathbf{W}^k$ ,  $\sigma$  is activation function, we use ReLU function.

Then, we normalized embedding vector  $\mathbf{h}_v^k$  as following Eq. (3):

$$\mathbf{h}_v^k = \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2 \quad (3)$$

Eventually, the node embedding vectors is shown in Eq. (4):

$$\mathbf{z}_v = \mathbf{h}_v^K \quad (4)$$

where  $K$  is the total number of information aggregation.  $\mathbf{z}_v$  is the embedding vector of the node  $v$ .

Given a node embedding vector, the probability of patient's syndrome is obtained through the calculation of the softmax, as shown in Eq. (5):

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}_v) \quad (5)$$

$\hat{\mathbf{y}}$  is the prediction probability vector. To train the model, the loss function adopts cross entropy, as shown in Eq. (6):

$$\mathcal{L}^P = -\frac{1}{|\mathcal{V}^P|_{train}} \sum_{i=1}^{|\mathcal{V}^P|_{train}} \mathbf{y}_i^P \cdot \log(\hat{\mathbf{y}}_i^P) \quad (6)$$

where  $|\mathcal{V}^P|$  is the total number of patients, and  $|\mathcal{V}^P|_{train}$  is the number of patients in training set,  $\mathbf{y}_i^P$  represents the label of the training set,  $\hat{\mathbf{y}}_i^P$  is the model output probability. The loss  $\mathcal{L}^P$  is optimized by stochastic gradient descent.

On the basis of the above, GCN model calculates  $\mathbf{h}_v^k$  in Eq. (7):

$$\mathbf{h}_v^k = \sigma(\hat{\mathbf{A}}\mathbf{h}_v^{k-1}\mathbf{W}^k) \quad (7)$$

where  $\hat{\mathbf{A}}$  is the normalized adjacency matrix.

RGCN adds different edge types to the GCN. We label the syndrome-symptom edge and the symptom-patient edge as different types. RGCN model calculates  $\mathbf{h}_v^k$  in Eq. (8):

$$\mathbf{h}_v^k = \sigma \left( \sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}_v^r} \frac{1}{c_{v,r}} \mathbf{W}_r^{k-1} \mathbf{h}_u^{k-1} + \mathbf{W}^{k-1} \mathbf{h}_v^{k-1} \right) \quad (8)$$

where  $\mathcal{R}$  is a collection of relational types.

### 3 Experiments and Results

#### 3.1 Dataset

The data of this study came from 539 case report forms of apoplexy patients provided by Heilongjiang University of Chinese Medicine. Among them, there are a total of 101 symptoms and 4 syndromes. 0 and 1 are used to mark whether a patient has a symptom or syndrome in the case report forms. The number distribution of each syndrome is shown in the following Table 1. Those syndromes refer to *Criteria of diagnosis and therapeutic effect of internal diseases and syndromes in traditional Chinese medicine (ZY/T 001.1-94)* and *Clinic terminology of traditional Chinese medical diagnosis and treatment*.

**Table 1** The number distribution of each syndrome

Syndromes	Wind and phlegm blocking collateral	Qi deficiency with blood stasis	Fu-organ excess due to phlegm and heat	Yin deficiency and wind stirring
Numbers	64	346	68	61

**Table 2** Results in the test set

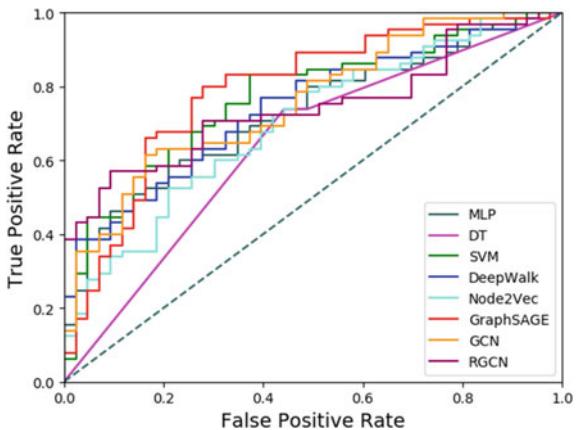
Model	Accuracy	Sensitivity	Specificity	AUC
MLP	0.67	0.60	0.77	0.73
DT	0.67	0.74	0.56	0.64
SVM	0.71	<b>0.75</b>	0.65	0.77
DeepWalk + LR	0.66	0.62	0.72	0.74
Node2Vec + LR	0.62	0.57	0.70	0.70
GraphSAGE	0.71	0.65	0.81	0.78
GCN	0.70	0.62	<b>0.84</b>	0.76
RGCN	<b>0.73</b>	0.69	0.79	<b>0.79</b>

### 3.2 Experimental Settings

We divided training set, validation set and test set according to 6:2:2. According to Table 1, the imbalance between these syndromes, so we use oversampling method to augment the small number of samples. This study classifies the syndrome of Qi deficiency with blood stasis and other syndromes, and sets the label of the Qi deficiency with blood stasis as 1 and the others as 0. Then we copy the sample of the 0 class once on the training set and the validation set. We use the accuracy, sensitivity, specificity and area under curve for measurement. Adam optimizer is used to calculate errors and update parameters, and the learning rate is 0.01. The dimension of  $\mathbf{h}_v^k$  is fixed at 16.

We choose the three most commonly used models in TCM syndrome differentiation, namely Multilayer Perceptron (MLP), Decision Tree (DT) and Support Vector Machine (SVM), these methods organize case report forms into structured data, 1 and 0 indicating a patient presence or absence of a certain symptom. Besides, we choose two classic graph embedding models DeepWalk [9] and Node2Vec [10], these two models embed the nodes in  $\mathcal{G}$  and take out the patient node vectors, then output the prediction syndrome through the logistic regression.

**Fig. 5** ROC curve for each model



### 3.3 Experimental Results and Discussion

The experimental results are shown in Table 2. Among the three models commonly used in TCM syndrome differentiation, SVM achieves the best results. DeepWalk and Node2Vec belong to graph embedding models, and their prediction results are close to classical classification models. The results of accuracy, specificity and AUC of the three popular graph neural network models are obviously better than the classical classification model and the classical graph embedding model. Moreover, RGCN has two indexes that are optimal among all models. The ROC curves are shown in Fig. 5, and those curves are consistent with results in Table 2.

## 4 Conclusion

In this study, the classical machine learning model, the graph embedding model and the current popular graph neural network model are used to classify the syndrome of apoplexy, and the data are obtained from the *Criteria* and case report forms.

The experimental results show that graph neural network models have obvious advantages. Moreover, RGCN introduces different edge types and obtains the best results in those models. To our knowledge, this study is the first time to apply the graph neural network model to TCM syndrome differentiation. The graph neural network models have surpassed the common models of TCM syndrome differentiation.

**Acknowledgements** This work was supported by the National Key Research and Development Program of China under Grant No. 2018YFC1707704.

## References

1. Zhao, C., Li, G.Z., Wang, C., Niu, J.: Advances in patient classification for traditional Chinese medicine: a machine learning perspective. *Evid. Based Complement. Altern. Med.* **2015**, 376716 (2015)
2. Xie, J., Li, Y., Wang, N., Xin, L., Fang, Y., Liu, J.: Feature selection and syndrome classification for rheumatoid arthritis patients with traditional Chinese medicine treatment. *Eur. J. Integr. Med.* **34**, 101059 (2020)
3. Yan, E., Song, J., Liu, C., Luan, J., Hong, W.: Comparison of support vector machine, back propagation neural network and extreme learning machine for syndrome element differentiation. *Artif. Intell. Rev.* **53**(4), 2453–2481 (2020)
4. Pang, H., Wei, S., Zhao, Y., He, L., Wang, J., Liu, B., Zhao, Y.: Effective attention-based network for syndrome differentiation of aids. *BMC Med. Inform. Decis. Mak.* **20**(1), 264–264 (2020)
5. Sun, Z., Yin, H., Chen, H., Chen, T., Cui, L., Yang, F.: Disease prediction via graph neural networks. *IEEE J. Biomed. Health Inform.* **25**(3), 818–826 (2020)
6. William, L.H., Rex, Y., Jure, L.: Inductive representation learning on large graphs. arXiv preprint [arXiv:1706.02216](https://arxiv.org/abs/1706.02216) (2017)
7. Thomas, N.K., Max, W.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations (2017)
8. Michael, S., Thomas, N.K., Peter, B., van den Berg, R., Ivan, T., Max, W.: Modeling relational data with graph convolutional networks. In: 15th International Conference on Extended Semantic Web Conference, ESWC 2018, pp. 593–607 (2018)
9. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. ACM (2014)
10. Aditya, G., Jure, L.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)

# Traditional Chinese Medicine Information Analysis Based on Multi-task Joint Learning Model



Chenyuan Hu<sup>✉</sup>, Zhuangzhi Yan, Jiehui Jiang, Shuoyan Zhang, and Tianyu Gu

**Abstract** In the field of traditional Chinese medicine (TCM) informatics, Chinese word segmentation and syndrome differentiation are two crucial analysis tasks. Owing to the ambiguity of the Chinese language and the peculiarities of syndrome differentiation, these tasks face huge challenges. Notably, from previous studies and investigations, these two tasks have a high correlation, which makes them fit the idea of multi-task joint learning (MTL). By sharing the underlying parameters and adding two different task loss functions, we proposed a novel MTL method to perform segmentation and classification of medical records in this research. Moreover, two classic deep neural network (Bidirectional LSTM (Bi-LSTM) and TextCNN) are fused into the MTL to conduct these two tasks simultaneously. As far as we know, our approach is the first attempt to combine these tasks with the idea of MTL. We used our proposed method to conduct a large number of comparative experiments. Through experimental comparison, it can be proved that our method is superior to other methods on both tasks. Therefore, this research can help to realize the modernization of TCM and the intelligent differentiation of TCM.

**Keywords** Traditional Chinese medicine · Chinese word segmentation · Syndrome differentiation · Multi-task joint learning · Deep learning

## 1 Introduction

Since traditional Chinese medicine (TCM) was incorporated into the latest global medical outline by World Health Organization (WHO), more and more scholars have begun to engage in research related to TCM [1, 2]. Evidence-based treatment is the basis of TCM, and accurate syndrome differentiation is significant to treatment. TCM syndrome differentiation refers to the use of TCM theories to analyze and summarize various disease data collected in the four diagnostic methods to find the key to the disease and point out the direction for clinical treatment. However,

---

C. Hu (✉) · Z. Yan · J. Jiang · S. Zhang · T. Gu  
Shanghai University, Shanghai 200444, China  
e-mail: [huchenyuan@shu.edu.cn](mailto:huchenyuan@shu.edu.cn)

a large amount of critical information about healthcare is buried in unstructured narratives, such as medical record, which makes computational analysis difficult. Moreover, environmental factors and empirical factors have a huge impact on the results of syndrome differentiation, which will lead to inaccurate and unstable diagnosis and treatment. Therefore, it is necessary and urgent to establish an objective and quantitative computer-aided syndrome differentiation method.

Specifically, automatic syndrome differentiation including the following two important technologies: Chinese word segmentation and text classification. In Chinese, words are the smallest language unit that can be used independently. Unlike English, Chinese words do not have clear separators between them, so word segmentation becomes a more important initial step. However, because the medical field has many professional vocabulary and there are ambiguities with modern Chinese, the task of TCM text segmentation is hard. The text classification task refers to automatically classifying text into several designated categories, so the intelligent syndrome differentiation of Chinese medicine can be abstracted as a text classification problem of the condition. Yin-yang is the general principle of the eight principles in TCM, and the realization of syndrome differentiation based on yin/yang can be a good foundation for subsequent medical judgments (such as treatment methods and formulas). Therefore, the task of Yin-Yang syndrome differentiation is a very fine classification task. Based on above, we try to use multi-task learning (MTL) and deep learning methods to solve these two challenging tasks of Chinese word segmentation and syndrome differentiation in our paper.

MTL is a very potential field in machine learning. Its goal is to use the useful information contained in multiple learning tasks to help learn a more accurate model for each task. The model can share information between different tasks, thereby improving the effect of the model. As reviewed in [3], they used the idea of MTL to integrate two important tongue characterization tasks into one model, and the effectiveness of their method was proved through experiments.

For those two crucial tasks in our study, worthy of affirmation, an excellent Chinese word segmentation result can retain correct, complete and important semantic information, which will promote to obtain better syndrome differentiation results. Similarly, the specified syndrome differentiation result can provide some additional features to assist in identifying certain specific semantic information, so as to obtain better word segmentation results. These two tasks are related rather than independent, which makes them consistent with the idea of MTL. According to the survey, MTL has demonstrated outstanding performance in many tasks, which inspired us to incorporate it into our research. Furthermore, our approach fuses a Bidirectional LSTM (Bi-LSTM) and a TextCNN into the MTL. To sum up, our research has the following three main contributions:

1. From the above, Chinese word segmentation and syndrome differentiation have a high correlation, which makes them fit the idea of MTL. As far as we know, our approach is the first attempt to combine these tasks with MTL. A large number of comparative experiments prove that our proposed method is superior to existing methods.

2. The model fuses two typical deep neural network into the MTL for Chinese word segmentation and TCM syndrome differentiation, including Bi-LSTM and TextCNN. And the model realizes end-to-end medical records analysis.
3. Each label is annotated and checked by TCM background personnel to ensure the reliability and accuracy of the data.

## 2 Related Works

### 2.1 Chinese Word Segmentation

Chinese word segmentation refers to the partitioning of a complete Chinese sentence into individual meaningful words, with the aim of facilitating the transformation of the words in the sequence into a computer-aware word vector in a subsequent task. Unlike English, Chinese words do not have clear separators between them, so Chinese word segmentation is designed to solve the problem of Chinese sentence segmentation.

Since the medical field has many professional vocabularies and there are ambiguities with modern Chinese, many research scholars have made a lot of efforts on the task of word segmentation in medical texts. Li et al. used dictionary and statistics-based word segmentation methods to segment Chinese medical record texts, and explored word segmentation methods suitable for medical texts [4]. Li et al. applied the capsule network (Capsule) to the word segmentation task of classical Chinese medicine books for the first time. In order to adapt the capsule structure to the sequence tagging task, a sliding capsule window was proposed, and the word segmentation accuracy on the public dataset reached 95% [5]. Xing et al. proposed a new Chinese word segmentation framework in the medical field, based on the Bidirectional LSTM-CRF model, using the multi-task learning framework of transfer learning, and using high-resource data to improve the word segmentation performance in the medical field [6]. Yuan et al. proposed an unsupervised Chinese word segmentation method based on a pre-trained BERT model, which was used for Chinese word segmentation and term discovery of electronic medical records, and achieved good performance [7].

### 2.2 Syndrome Differentiation

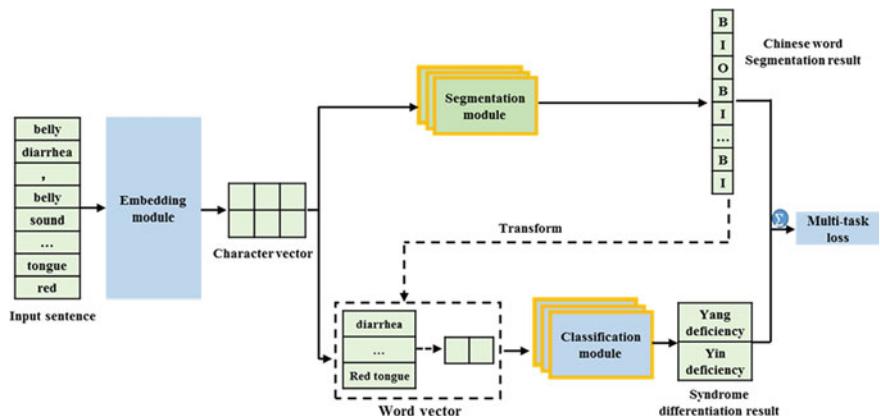
Medical records include the symptoms of patients and the corresponding syndrome differentiation and treatment methods, which are of special significance for the inheritance of the experience of famous old doctors of TCM. To better generalize the experience of TCM experts, we modeled syndrome differentiation for Yin/yang deficiency in the form of medical record text classification [8, 9].

With the rapid development of machine learning and deep learning algorithms, more and more text classification techniques have been widely used in modern TCM syndrome differentiation research. Li uses the subject-related weighting model to classify TCM medical records [10]; The support vector machine was used to study the syndrome differentiation of patients with depression, and a high classification accuracy was obtained [11]. Zhao et al. proposed to explore the relation between syndromes for viral hepatitis by using manifold ranking (MR) [12]. These studies mainly used machine learning algorithms. Besides, there are also many researches on TCM syndrome differentiation using deep learning algorithms. For example, a deep belief network is used to construct a diagnosis model of TCM chronic gastritis syndromes in [13]. Zhu et al. propose a deep learning algorithm for TCM damp-heat syndrome differentiation [14]. Hu et al. use two neural network models to realize the differentiation of Yin and Yang in TCM [15].

These methods described above have achieved good performance. However, there are still some disadvantages: (1) Some researches on word segmentation of TCM texts require the construction of TCM corpus, which is labor intensive; (2) Some syndrome differentiation studies ignore the inner connection between these two tasks.

### 3 Method

In order to realize end-to-end medical records analysis, improving the performance of Chinese word segmentation and syndrome differentiation. We propose a novel model that combines the idea of MTL, which can conduct these tasks simultaneously. As shown in Fig. 1, the segmentation module is used for Chinese word segmentation, and syndrome differentiation is carried out by the classification module. Besides, the role of the common embedding module is to provide shared information for these



**Fig. 1** An overview of our framework

tasks. In this section, firstly, the three modules are explained separately, and then the joint loss function is introduced.

### 3.1 Embedding Module

Defined in pytorch, the embedding module is a simple lookup table for storing fixed dictionaries and size embedding. This module mainly relies on indices to retrieve word embedding. The input of the module is the index list, and the output is the corresponding word embedding. Through this module, we can get the vector representation of the sentence.

### 3.2 Segmentation Module

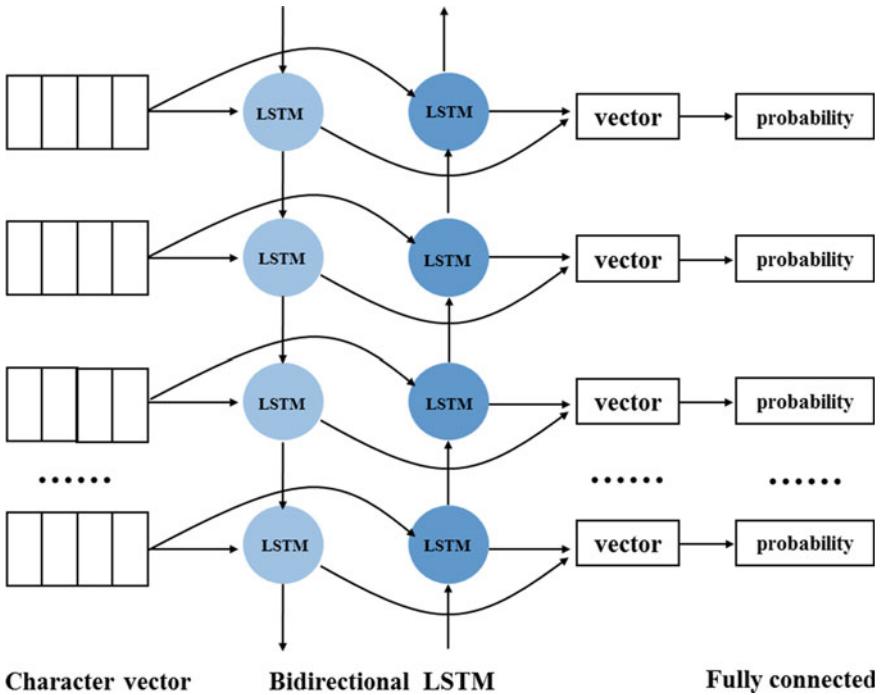
In order to transform the Chinese word segmentation task into a sequence tagging problem, a label is assigned to each character. There are three types of labels:  $B$ ,  $I$ , and  $O$ , which correspond to the beginning, middle and end of words, and single-word characters, respectively. Formulate the problem, given a sequence with  $n$  characters  $X = \{x_1, \dots, x_n\}$ , the purpose of the Chinese word segmentation is to find the mapping from  $X$  to  $Y^* = \{y_1^*, \dots, y_n^*\}$ :

$$Y^* = \arg \max_{Y \in \mathcal{L}^n} p(Y|X) \quad (1)$$

where,  $\mathcal{L} = \{B, I, O\}$ .

The LSTM unit can learn long-term dependencies without retaining redundant context information. And it has been proven to have a good performance on sequence tagging tasks, and it is now widely used in natural language processing tasks. The Bi-LSTM is composed of LSTM units, and has two parallel levels and propagates in two directions, which can utilize past and future input features. Therefore, this research uses a Bi-LSTM models to perform Chinese word segmentation.

The module structure is shown in Fig. 2. The vector from embedding layer is fed into the Bi-LSTM network to obtain the past and future spliced feature vectors, then obtain the probability of each label through the fully connected layer, and obtain the tag with the maximum probability. By this, we can obtain the sequence tagging results.



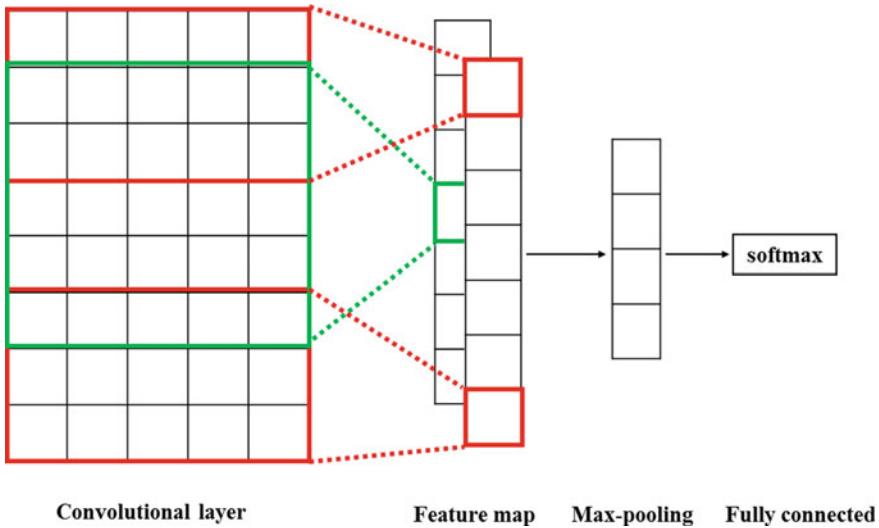
**Fig. 2** Structure of segmentation module

### 3.3 Classification Module

Convolutional neural networks (CNN) are widely used in tongue color classification [16], cracked tongue diagnosis and pulse signal classification [17] in TCM. In our research, we use the CNN model to classify TCM medical record texts, so it is also called TextCNN. Its system structure is shown in Fig. 3, for the obtained character vector, in the convolutional layer, multiple convolution kernel sizes ([3, 5]) are used to convolve the embedded vectors. Then, the vectors are passed the max-pooling layer to capture the most salient features. Finally, the classification results are obtained through the fully connected layer.

### 3.4 Joint Loss Function

In essence, the loss functions of the Chinese word segmentation task and the syndrome differentiation task are both cross-entropy, but the difference is that the Chinese word segmentation task uses masked cross-entropy to avoid the influence of padding characters. As reviewed in [18], in order to optimize all the parameters



**Fig. 3** Structure of classification module

involved in multi-task, the loss function of the multi-task model is defined as the weighted sum of the loss functions of different tasks. These weights are called as hyperparameters. Based on the above, our loss function is defined as Eq. (2). Among them,  $L_0, r_0$  are the loss function and the corresponding weight of the Chinese word segmentation task. Similarly,  $L_1, r_1$  are the loss function and the corresponding weight of the TCM syndrome differentiation task. Our training strategy is to minimize the *Loss*.

$$\text{Loss} = r_0 * L_0 + r_1 * L_1 \quad (2)$$

## 4 Experiments

In order to prove the effectiveness of the proposed model, we conducted a lot of comparative experiments. For the convenience of expression, we use JCS to represent our proposed joint model, that is, the first letter of the Joint Chinese word segmentation and Syndrome differentiation.

## 4.1 Dataset

The “Essences of Modern Chinese Medical Records of Modern Chinese Medicine” series mainly contains medical records of many famous doctors of TCM in the country [19]. Each medical record is personally selected by the famous doctor himself, and its content is true and reliable. Therefore, we use these medical records in the fifth and sixth episodes of this series as the original dataset. In order to ensure the validity and accuracy of our research, it is necessary to preprocess the medical record texts, including text extraction, and invite people with a background in TCM education to conduct an annotation and inspection [20, 21]. After the above processing, a total of 1209 medical records were obtained, including 643 cases of Yang deficiency syndrome and 566 cases of Yin deficiency syndrome.

## 4.2 Training Details

In our study, we split the dataset into training (60%), validation (10%), test (30%) datasets randomly. The optimizer is adaptive moment estimation (Adam), the epochs are 50, and the learning rate is 0.001. During the training process, when the loss of the validation dataset is minimal, saving the model for testing. That is, saving the optimal model for testing. And optimizing hyperparameters in the loss function by random search method, the list is [0.1, 0.2, 0.3, 0.4, 0.5], and the best experimental results can be obtained when the values of  $r_0$ ,  $r_1$  are set to 0.4 and 0.4.

## 4.3 Chinese Word Segmentation Experiments

This part is the experiments of Chinese word segmentation. We compare the JCS model with some basic models with the same parameter settings. And in order to meet the needs of the project team, the evaluation criteria included accuracy, specificity and sensitivity.

**Baselines.** We compare our method with several LSTM-based models, including LSTM, Bi-LSTM, Bi-GRU. LSTM is a variant of the RNN model [22]. It mainly uses three gates: forgetting gate, input gate, and output gate to achieve the role of information transmission [23]. Gated Recurrent Unit (GRU) replaces the forget gate and input gate in LSTM with update gate [24]. The Bidirectional LSTM/GRU network is similar in structure to the LSTM/GRU network. The difference is that they have two parallel levels and propagates in two directions.

**Results analysis.** The results of different models are shown in Table 1. Compared with the best baseline model, the JCS model has improved by 2.65%/3.45%/8.11%

**Table 1** Comparison results between JCS and different models

Methods	Accuracy (%)	Specificity (%)	Sensitivity (%)
LSTM	87.52	89.05	79.25
Bi-LSTM	85.36	88.68	78.52
Bi-GRU	85.57	88.68	77.78
JCS	<b>90.17</b>	<b>92.50</b>	<b>87.36</b>

in accuracy/specificity/sensitivity. Therefore, the above comparison can prove that our method has better performance on the task of Chinese word segmentation.

#### 4.4 Syndrome Differentiation Experiments

This part is the experiments of syndrome differentiation. We conducted a lot of comparative experiments between the JCS model and existing methods. And in order to meet the needs of the project team, the evaluation criteria included accuracy, specificity and sensitivity. Besides, we also use running time as a performance evaluation indicator.

**Baselines.** In contrast to our JSC model, there are two situations in the existing method. (1) one-stage method. only have the syndrome differentiation, ignore the impact of the Chinese word segmentation; (2) two-stage methods. The first stage is Chinese word segmentation, and the second stage uses text classification methods to classify syndromes. In order to ensure the comparability and fairness of the experiment, for the existing methods, on the one hand, we use the TextCNN model for the syndrome differentiation; on the other hand, in the first step, we employ the Bi-LSTM model for Chinese word segmentation, and then the result is used as the input of the second step for syndrome classification. Classic text classification methods include traditional machine learning methods and deep learning methods. Therefore, the traditional method is support vector machine (SVM), and the commonly used TextCNN and TextRNN are adopted as neural network methods in our study.

**Results analysis.** The comparison results of this task are shown in Table 2. First of all, as shown in the first and fourth lines, the performance of the two-stage method is better than the performance of only one stage, which proves the necessity of Chinese word segmentation in the field of Chinese medicine. Then, compared with the relatively best baseline model, the JCS model has improved in most indicators, with an increase of 3.71% and 11.79% in accuracy and specificity, respectively. Besides, we can see that these baseline models are based on the same datasets where the number of two categories is close to 1:1, since the difference between specificity and sensitivity is about 10%. Furthermore, our method only needs one feature extraction operation, the running time of our JCS model is only 6 min, which is faster than those two-stages methods, since these two-stages methods need to conduct feature extraction twice.

**Table 2** Comparison results between JCS and existing methods

First stage	Second stage	Joint/none-joint	Accuracy (%)	Specificity (%)	Sensitivity (%)	Time (min)
–	TextCNN	None-joint	73.48	70.73	87.86	3.5
Bi-LSTM	SVM	None-joint	75.21	70.99	78.61	14.9
Bi-LSTM	TextRNN	None-joint	69.06	86.71	52.91	17.5
Bi-LSTM	TextCNN	None-joint	76.24	83.24	69.84	15.9
–	–	Joint(JCS)	<b>79.95</b>	<b>78.97</b>	<b>81.33</b>	<b>6</b>

Through the above description, our method has a more excellent performance in the task of syndrome differentiation compared with these existing methods.

## 5 Conclusion

As discussed above, Chinese word segmentation and syndrome differentiation have a high correlation, which makes them fit the idea of MTL. Combining the ideas of MTL, we propose a novel method to perform word segmentation and classification of medical records. Moreover, two classic deep neural network (Bi-LSTM and TextCNN) are merged into the MTL to conduct these two tasks simultaneously. As far as we know, our approach is the first attempt to combine these tasks with MTL. We used the proposed method to conduct a large number of comparative experiments. Through experimental comparison, it can be proved that our method is superior to other methods on both tasks. However, there are still the following two works to improve the proposed method, since the proposed method still does not achieve outstanding performance.

1. Since deep learning models always achieve better performance in larger datasets, more medical records will be obtained in the future.
2. We will use more advanced models to improve the performance of the joint model in the future.

**Acknowledgements** This work was supported by the National Key Research and Development Program of China under Grant No 2018YFC1707704.

## References

1. Cyranoski, D.: The big push for Chinese medicine for the first time, the World Health Organization will recognize traditional medicine in its influential global medical compendium. *Nature* **561**(7724), 448–450 (2018)

2. Seung-Hoon, C.: A milestone in codifying the wisdom of traditional oriental medicine: TCM, Kampo, TKM, TVM-WHO international standard terminologies on traditional medicine in the western Pacific region. *ECAM* **7**(3), 303–305 (2016)
3. Qiang, X.: Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network. *IEEE J. Biomed. Health Inform.* **24**(9), 2481–2489 (2020)
4. Goulei, L.: Research on segmentation of Chinese text in medical record. *Chin. J. Biomed. Eng.* **35**(4), 477–481 (2016)
5. Si, L.: Capsules based Chinese word segmentation for ancient Chinese medical books. *IEEE Access* **6**, 3619–3630 (2018)
6. Junjie, X.: Adaptive multi-task transfer learning for Chinese word segmentation in medical text. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, pp. 3619–3630 (2018)
7. Zheng, Y.: Unsupervised multi-granular Chinese word segmentation and term discovery via graph partition. *J. Biomed. Inf.* **110** (2020)
8. Miao, J.: Syndrome differentiation in modern research of traditional Chinese medicine. *J. Ethnopharmacol.* **140**(3), 634–642 (2012)
9. Liang, Y.: Traditional Chinese medicine clinical records classification using knowledge-powered document embedding. In: IEEE International Conference on Bioinformatics and Biomedicine, pp. 1926–1928. IEEE Computer SOC, USA (2016)
10. Yiming, L.: Cross-domain learning based traditional Chinese medicine medical record classification. In: 10th International Conference on Intelligent Systems and Knowledge Engineering, pp. 335–340. IEEE, USA (2015)
11. Jianglong, S.: A network-based approach to investigate the pattern of syndrome in depression. *Evidence-based Complement. Altern. Med.* **2015** (2015)
12. Yufei, Z.: Syndrome classification based on manifold ranking for viral hepatitis. *Chin. J. Integr. Med.* **20**(5), 394–399 (2014)
13. Guoping, L.: Deep learning based syndrome diagnosis of chronic gastritis. *Comput. Mat. Methods Med.* **2014** (2014)
14. Wei, Z.: A study of damp-heat syndrome classification using Word2vec and TF-IDF. In: IEEE International Conference on Bioinformatics and Biomedicine, pp. 1415–1420. IEEE Computer SOC, USA (2016)
15. Qinan, H.: End-to-end syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine. *Comput. Methods Programs Biomed.* **174**, 9–15 (2019)
16. Shiru, Z.: Human pulse recognition based on convolutional neural networks. In: 2016 International Symposium on Computer, Consumer and Control. IEEE, USA (2016)
17. Jun, H.: Classification of tongue color based on CNN. In: 2nd IEEE International Conference on Big Data Analysis. IEEE, USA (2017)
18. Yu, Z.: An overview of multi-task learning. *Natl. Sci. Rev.* **5**(1), 30–43 (2018)
19. Yongyan, W.: The Essence of Medical Records of Modern Famous Chinese Medicine. People's Medical Publishing House, Beijing (2010)
20. Lu, F.: Discussion on the standard of word segmentation of ancient Chinese medicine books: taking the medical books of Qing dynasty as an example. *China J. Tradit. Chin. Med. Pharm.* **33**(10), 4700–4705 (2018)
21. Candong, L.: Diagnostics of Traditional Chinese Medicine. China Press of Traditional Chinese Medicine, Beijing (2016)
22. Zhiheng, H.: Bidirectional LSTM-CRF models for sequence tagging. In: Computer Science (2015)
23. Yushi, Y.: Bi-directional LSTM recurrent neural network for Chinese word segmentation. In: 23rd International Conference on Neural Information Processing, pp. 345–353. Springer, Switzerland (2016)
24. Cho, K.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Computer Science (2014)

# TongueCaps: A Model for the Multiclassification of Tongue Color



Jinghong Ni , Zhuangzhi Yan , and Jiehui Jiang

**Abstract** Tongue colors can reflect physiological information in the human body and assist Chinese medicine doctors in completing syndrome differentiation and treatment, so the classification of human tongue color is of great significance. However, for the three more common tongue colors, light red, red and deep red, their color tone is all red, in addition, because the color rendering of the tongue image is affected by the color temperature of the ambient light, camera equipment and color rendering equipment, it is difficult to classify the color of tongue. In our work, we propose TongueCaps, a model based on the combination of residual block and capsule, to realize end-to-end computer-aided tongue color classification. This paper carries out classification experiments on three categories of tongue color with RGB images, light red tongue (382 cases), red tongue (312 cases), and deep red tongue (104 cases). Furthermore, we tested TongueCaps in comparison with VGG16, ResNet18. The model effect was evaluated by accuracy, sensitivity, specificity. Compared with the results, TongueCaps showed the best performance.

**Keywords** Tongue color · Residual block · Capsule · Artificial intelligence

## 1 Introduction

Tongue diagnosis is an important content of traditional Chinese medicine inspection, and it is an important observation method in clinical practice. Doctors understand the physiological functions and pathological changes of the human body by observing the characteristics of the tongue, such as color, texture, tooth marks, and tenderness, and realize the dialectical differentiation of the human body [1]. Recently, with the gradual development of computer image processing technology, more and more researches are devoted to the use of computer image processing technology to processing and analysis of tongue diagnosis images, such as automatic tongue image segmentation [1–3], cracked tongue recognition and classification [4–6], tooth

---

J. Ni · Z. Yan · J. Jiang  
Shanghai University, Shanghai 200444, China  
e-mail: [njh123@shu.edu.cn](mailto:njh123@shu.edu.cn)

marked tongue recognition [7–9], tongue image color and texture feature analysis [10, 11], using tongue image features to identify diabetic patients [12], etc. Tongue diagnosis with the help of computer image processing technology can effectively avoid the subjectivity, ambiguity and other shortcomings of traditional tongue diagnosis, and lead traditional tongue diagnosis to the development of digitization, objectivity, and intelligence, and realize computer-aided diagnosis.

Tongue color is an important content in tongue diagnosis. Tongue color, that is, the color of the tongue texture. Different tongue colors can reflect the different physiological information of the human body. Light red tongue, red tongue, and deep red tongue are three common tongue colors. When they are classified, they are difficult to classify because of the following reasons: (1) Their color tones are all red. (2) The tongue texture and tongue coating are interlaced on the tongue, and the color of tongue coating has an influence on the classification of tongue color. (3) The amount of sample data is small and the resolution is low.

At present, TCM doctors will generally via visual observation and clinical experience to identify tongue color types. This method is easily affected by factors such as ambient light, doctors' color sensitivity, and doctors' clinical experience. Therefore, the tongue color recognition results of the same patient under different doctors or different environments may be different, resulting the tongue color recognition results to be subjective, lacking objectivity, quantification, and standardization.

## 2 Previous Works

In the early days, most researchers select a specific area of the tongue as the object of tongue color research, and then quantitatively analyze the different types of tongue colors based on the principle of colorimetry, and then used statistical analysis methods to analyze them, trying to find the differences in the color values of different types of tongue colors, providing an objective basis to classify the tongue colors. [13, 14] use the RGB color space to quantitatively analyze the chromaticity value of tongue color, although there is a certain difference in tongue color chromaticity value between the two studies, it is found that the B chromaticity value of the purple tongue is the highest and the R chromaticity value is the lowest. There are also some studies that statistically analyze diseases and tongue color values. Weng [15] use RGB color space quantitatively analysis the tongue color of blood stasis syndrome and non-blood stasis syndrome. Xu [16] found that the B chromaticity ratio of the tongue of breast cancer patient is higher than R chromaticity and G chromaticity. Chen [17] and Fu [18] quantitatively analyze RGB component values of the tongue color of patients with rheumatoid arthritis. Yang [19] Yang Xinyu summarized the past 22 studies on the quantitative analysis of tongue color based on the CIE Lab color space, and found that from light white tongue to light red tongue, red tongue, and deep red tongue, regular changes in chromaticity values were observed, which showed a gradual decrease in the L value. Kawanabe [20] separated tongue body and coating by clustering, calculated the average value of each component of LAB for pixels

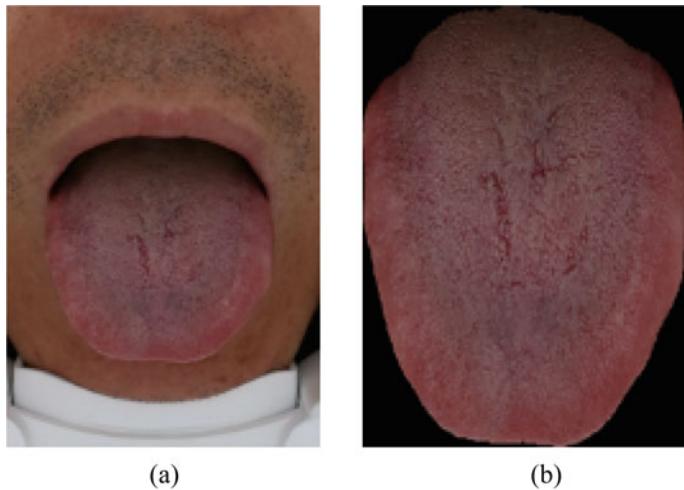
belonging to tongue body, and found there is a statistical difference between light red tongue, red tongue and deep red tongue on the L and B components. However, due to the different data and methods used in different studies, there is a large gap between the chromaticity values of each type of tongue color, and it is impossible to select one of them as the standard for tongue color classification.

In the later period, most researches apply the method of pattern recognition to classify tongue color. Li [21] based on the hyperspectral tongue image, use the reflection spectrum angle mapping algorithm to match the tongue color spectrum with the known color spectrum to achieve tongue color classification. Wang [22] used 90,000 tongue images taken under standard light sources to define the tongue color gamut, and looking for 12 representative colors in the tongue color gamut to extract color features for classification. Kamarudin [23] use support vector machines to perform two-stage classification. Tang [24] proposed a tongue image classification method based on multi-task convolutional neural network, trying to realize the simultaneous recognition of multiple labels such as tongue color, coating color, cracks and tooth marks. Jiao [25] used an unsupervised learning method, K-Means, to cluster the four tongue colors. Zhang [26] used data that manually selected by experts, proposed a support vector machine-based tongue body color and coating color recognition method for patients with acne.

At present, the research methods of tongue color are mostly biased towards traditional machine learning, and the research methods are quite different and lack certain universal applicability. In addition, deep learning has shown good superiority in other applications in the field of image processing, so this article will use the deep learning model to conduct experiments. At the same time, the difficulty of acquiring tongue images of different tongue colors is different, which makes the amount of data less. Therefore, in this research, we propose a model that combines the residual module and the capsule network to classify the red tongue, red tongue, and deep red tongue.

### 3 Materials

The raw image data used in the experiment in this study were all acquired by the tongue image acquisition equipment of Tianjin Huiyigu Technology Co., Ltd., and the acquisition process was the face of the subject is facing the tongue image acquisition device, and the tongue is fully squeezed out. Figure 1a is an example of raw tongue image data. The raw image is a 24-bit RGB image, and the image size includes two types, which are  $1640 \times 2460$  and  $1480 \times 2220$ . The tongue image is automatically segmented from the raw tongue image by using the tongue image acquisition instrument, as shown in Fig. 1b, which separates the tongue body from other tissues on the image for subsequent research. The purpose of image tongue segmentation is to remove the influence of lip color, face color, and tongue coating color, etc. on tongue color recognition.



**Fig. 1** **a** Raw image. **b** Image after segmentation

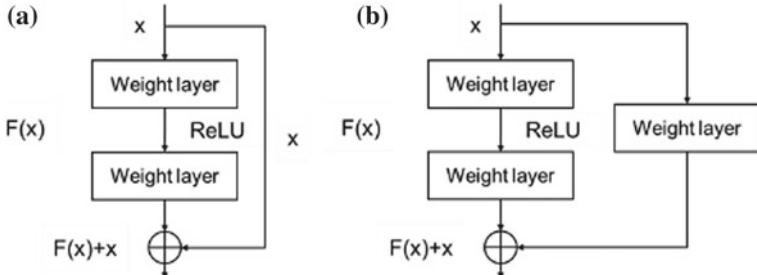
**Table 1** Sample image data statistics

Tongue color type	Light red	Red	Deep red
Image number	382	312	104

In total, 798 cases of tongue images are used for experiment, the statistics of experimental sample data are shown in Table 1, including 382 cases of light red tongue, 312 cases of red tongue, and 104 cases of deep red tongue.

## 4 Methods

The pooling layer used by Traditional Convolutional Neural Network will lose a lot of information about spatial location, while CapsNet discards the pooling layer in order to keep as much spatial location information as possible. In addition, the convolutional layer of CapsNet has only one layer, the feature extraction ability is weak, and it is unable to fully learn the features of the image, so the method in this paper improves the capsule network in order to obtain a better performance model. This section gives a detailed introduction to the framework of the model used in our work, including the layers contained in the model, the forward propagation process of the model, and the principle updating the model parameter.



**Fig. 2** **a** Residual block a. **b** Residual block b

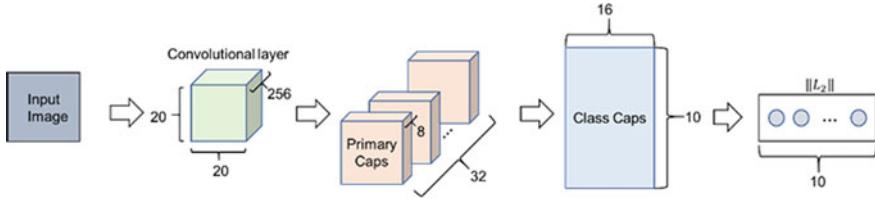
#### 4.1 Residual Block

ResNet [27] can effectively solve the problem of gradient disappearance caused by the increase of the number of layers in traditional convolutional neural networks. The main structural feature of ResNet is the Residual Block. As shown in Fig. 2, the jump connection structure is adopted. The original input and the feature map obtained through the weight layer are added and sent to the back layer of the network, which can improve the efficiency of model training and make the number of model layer is not restricted by the problem of gradient disappearance, which solves the over-fitting problem in the model training process and improves the generalization ability of the model. According to whether the size of the output feature map of the weight layer is consistent with the size of the original input  $x$ , the residual block generally has two types, as shown in Fig. 2, where (a) represents the size of the feature map output by the weight layer and the original input  $x$  is the same, (b) indicates that the size of the feature map output by the weight layer is inconsistent with the size of the original input  $x$ , and the original input  $x$  needs to be convolved and then spliced with the feature map.

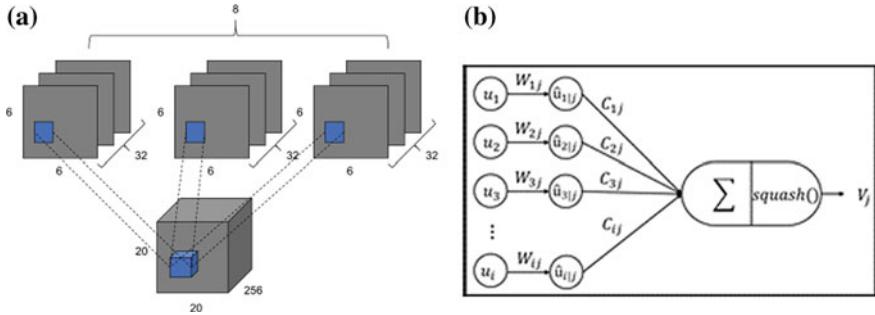
#### 4.2 Capsule Network

The CapsNet (Capsule Network) is a network model proposed by Professor Hinton [28]. This network takes into account the relative position, angle and other information that CNN lacks, so compared to CNN, it can better recognize images of the same type of object in different perspectives, and is closer to the way of thinking of the human brain. For small sample training sets, achieve the effect of analogy. The network mainly composed of three parts, as shown in Fig. 3, Convolutional layer, Primary Caps, and Class Caps.

**Primary Caps and Class Caps.** Convolutional layer output are feature maps. After the output of the convolutional layer, Primary Caps turns the input feature maps into vector as output. Primary Caps consists of two parts of operation, convolution



**Fig. 3** Capsule network



**Fig. 4** **a** Convolution of primary caps. **b** Capsule

and capsule. For example, the size of the input feature maps is  $256 \times 20 \times 20$  (channels, width, height), the Primary Caps has 32 units, every units has 8 channels convolutional layer and every convolutional layer with  $9 \times 9$  kernel size and strides 2. Then the output of every channels are feature maps with size of  $6 \times 6 \times 32$ , output of 8 channels are  $6 \times 6 \times 8 \times 32$ , as shown in Fig. 4a. Every channel converts feature maps to a vector, so the output of Primary Caps are 8 vectors with dim of each vector is  $6 \times 6 \times 32 = 1152$ .

Then, 8 vectors are fed into capsule in end of Primary Caps. The output of each capsule nerve represents the probability that an entity is contained in the image, that is, the output of a capsule is a vector. Capsule nerves are similar to traditional neurons, but they are different from traditional neurons. The forward propagation of capsule as shown in Fig. 4b,  $u_i$  is the output of capsule  $i$  in layer  $l$ ,  $u_j$  is the output of capsule  $j$  in layer  $(l + 1)$ . First transform the input, as shown in Formula (1), multiply the  $u_i$  by the weight  $w_{ij}$  to get the prediction vector  $\hat{u}_{j|i}$ .

$$\hat{u}_{j|i} = W_{ij}u_i \quad (1)$$

Then perform a weighted sum on the prediction vector according to Formula (2) to get  $s_j$ , among them,  $c_{ij}$  is the weight coefficient, which is updated by the dynamic routing algorithm.

**Table 2** Procedure of dynamic routing algorithm

---

**Procedure1** Dynamic routing algorithm

---

**procedure** Dynamic routing ( $\hat{u}_{j|i}, r, l$ )

for all capsule  $i$  in layer  $l$  and capsule  $j$  in layer  $(l + 1)$ :

$b_{ij} \leftarrow 0$ .

**for**  $r$  iteration **do**

for all capsule  $i$  in layer  $l$ :  $c_i \leftarrow softmax(b_i)$

for all capsule  $j$  in layer  $(l + 1)$ :  $s_j \leftarrow \sum_i c_{ij} \hat{u}_{j|i}$

for all capsule  $j$  in layer  $(l + 1)$ :  $V_j \leftarrow squash(s_j)$

for all capsule  $i$  in layer  $l$  and capsule  $j$  in layer  $(l + 1)$ :

$b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} V_j$

**return**  $V_j$

---

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \quad (2)$$

Finally, use Formula (3) to operate nonlinear activation, ensure the length of output  $V_j$  in the interval  $[0, 1]$ .

$$V_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \quad (3)$$

**Dynamic routing algorithm.** When the capsule forward propagation was introduced in the previous section, it was mentioned that  $c_{ij}$  is updated through dynamic routing algorithm. This algorithm is introduced in Table 2, among them,  $w_{ij}$  is affine transform matrix, is updated by back propagation,  $b_{ij}$  in the table is a parameter that needs to be initialized, then use Formula (4) calculate  $c_{ij}$ ,

$$softmax(b_i) : c_{ij} = \frac{\exp(b_{ij})}{\sum_j \exp(b_{ij})} \quad (4)$$

In fact, the direction of the update of the weight  $c_{ij}$  is to give a large weight to the output vector of the capsule neuron in the layer  $l$  that has a large contribution to the final recognition.

**Loss Function.** The capsule network uses the vector length to represent the probability of the existence of each entity. When the entity appears in the image, it is hoped that the loss will be small, and when the entity does not exist, it is hoped that the loss will be large, so the marginal loss is used in Formula (5),

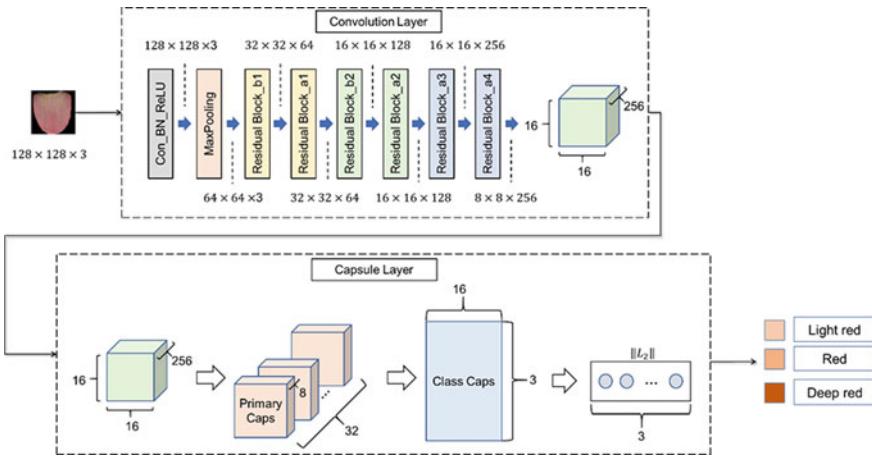
$$L_k = T_k \max(0, m^+ - \|V_k\|)^2 + \lambda(1 - T_k) \max(0, \|V_k\| - m^-)^2 \quad (5)$$

Among them,  $T_k$  is the classification indicator function (class k exists, the value is 1, otherwise it is 0);  $V_k$  is the output vector of the net;  $m^+$  is used to punish false positives, the value is 0.9;  $m^-$  is used to false negatives, the value is 0.1;  $\lambda$  is the proportional coefficient, adjust the proportion of the two punitive, the value is 0.5.

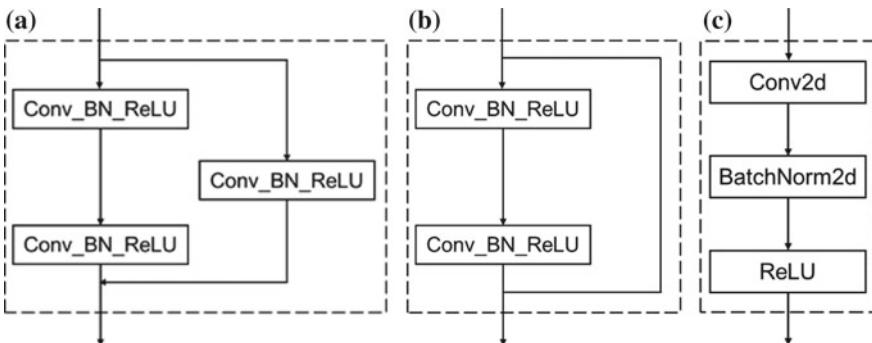
### 4.3 TongueCaps

The framework of TongueCaps is shown in Fig. 5, including two parts, convolution layer and capsule layer.

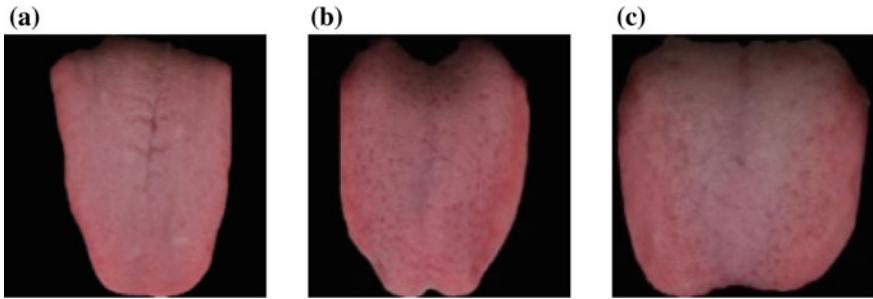
The convolution layer use residual block. Residual Block a and b as shown in Fig. 6.



**Fig. 5** The framework of TongueCaps



**Fig. 6** **a** Residual block a and **b** residual block b, **c** the weight layer Conv\_BN\_ReLU



**Fig. 7** **a** Light red, **b** red, **c** deep red

## 5 Experiment and Results

### 5.1 Data Preprocessing and Data Division

The size of the tongue image data is different. For the convenience of experiment, the image size is fixed to a uniform size of  $128 \times 128$  by using equal scaling, as shown in Fig. 7.

Each type of tongue color in the data set is divided into training set and test set at 8:2. Due to the small number of data samples, the training data is expanded before the experiment. Considering that the method of expansion cannot affect the color information of the tongue image, the methods of horizontal movement, rotation, and vertical movement are adopted for expansion.

### 5.2 Set Parameters of Training

The optimizer for model training selects Adam, and the learning rate selects 0.001. The hyperparameter batchsize is selected by the grid search method. The selection interval of the batch size is [10, 50], and the stride is 10. After each setting determines the batchsize, use fivefold cross-validation to calculate the average accuracy of the 5 models on the validation set, and finally select the hyperparameter value corresponding to the model with the largest average accuracy to train the entire training set. Get the final model parameters, and check the classification effect on the test set.

### 5.3 Results

Our work use the accuracy, sensitivity and specificity to measure the performance. Calculation of three assessment criteria as shown in Formulas (6–8),

**Table 3** Results compared with other networks

	Accuracy	Sensitivity	Specificity
VGG16	$0.4402 \pm 0.0541$	$0.3275 \pm 0.0101$	$0.6626 \pm 0.0070$
ResNet18	$0.6805 \pm 0.0714$	$0.7056 \pm 0.0486$	$0.8135 \pm 0.0316$
TongueCaps	$0.7308 \pm 0.0396$	$0.7671 \pm 0.0254$	$0.8479 \pm 0.0165$

**Table 4** The size of TongueCaps is much smaller than the other two models

Models	Model size, MB
VGG16	384.74
ResNet18	134.29
TongueCaps	8.09

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

where TP is the number of positive samples correctly predicted by the model, TN is the number of negative samples correctly predicted by the model, FP is the number of positive samples incorrectly predicted by the model, FN is the number of negative samples incorrectly predicted by the model.

The results of our work and compared with other networks as shown in Table 3, it can be found that the performance of TongueCaps are higher than VGG16, ResNet18 on Accuracy, Sensitivity, Specificity. Furthermore the size of model as shown in Table 4.

From the results, it can be seen, compared to the other two networks, TongueCaps has smaller model parameters but also higher accuracy, sensitivity, and specificity.

## 6 Conclusion

In this paper, we proposed TongueCaps to classify light red tongue, red tongue and deep red tongue. The experiment data are RGB images. The framework of TongueCaps combine the advantages of both Residual block and Capsule network, can fully learn the characteristics related to tongue color, avoiding the adverse effects of tongue coating and other factors on tongue color recognition. Compared with VGG16 and ResNet18, our model has the best performance in assessment criteria, accuracy, sensitivity, and specificity. In addition, TongueCaps has smaller model size

than VGG16 and ResNet18. This lays the groundwork for a new method to classify tongue color more simply.

**Acknowledgements** This work was supported by the National Key Research and Development Program of China under Grant No 2018YFC1707704.

## References

1. Li, Z., Yu, Z., Liu, W., Xu, Y., et al.: Tongue image segmentation via color decomposition and thresholding. *Concurrency Comput. Pract. Experience* **31**(23) (2019)
2. Zhou, C., Fan, H., Li, Z.: Tonguenet: accurate localization and segmentation for tongue images using deep neural networks. *IEEE Access* **148779–148789** (2019)
3. Zhou, J., Zhang, Q., Zhang, B., et al.: TongueNet: a precise and fast tongue segmentation system using U-Net with a morphological processing layer. *Appl. Sci. Basel* **9**(15) (2019)
4. Huo, C.M., Zheng, H., Su, H.Y., et al.: Tongue shape classification integrating image preprocessing and convolution neural network. In: Proceedings 2017 2ND ASIA-PACIFIC Conference on Intelligent Robot Systems, pp. 42–46 (2017)
5. Sudarshan, R., Vijayabal, G.S., Samata, Y., et al.: Newer classification system for fissured tongue: an epidemiological approach. *J. Trop. Med.* <https://doi.org/10.1155/2015/262079> (2015)
6. Zhang, H.K., Hu, Y.Y., Li, X., et al.: Computer identification and quantification of fissured tongue diagnosis. In: Proceedings 2018 IEEE International Conference on Bioinformatics and Biomedicine, pp. 1953–1958 (2018)
7. Wang, X., Liu, J., Wu, C., et al.: Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark. *Comput. Struct. Biotechnol. J.* **18**, 973–980 (2020)
8. Sun, Y., Dai, S., Li, J., et al.: Tooth-marked tongue recognition using gradient-weighted class activation maps. *Future Internet* **11**(2) (2019)
9. Tang, W., Gao, Y., Liu, L., et al.: Automatic recognition of tooth-marked tongue based on tongue region detection and tongue landmark detection via deep learning. *IEEE Access* **8**, 153470–153478 (2020)
10. Luo, Y., Wang, Z.F.: A texture feature extraction of tongue image based on weighted binary method. *Inf. Technol.* **7**, 49–51,55 (2017)
11. Zhu, W.N., Zhou, C.L., Xu, D., et al.: Application of multi feature image retrieval technology based on color texture in tongue diagnosis of traditional Chinese medicine. *Chin. J. Image Graphics* **8**, 57–63,129 (2005)
12. Wu, L., Luo, X., Xu, Y.: Using convolutional neural network for diabetes mellitus diagnosis based on tongue images. *J. Eng. JoE* **13**, 635–638 (2020)
13. Zhang, S.H., Guo, A.Y., Liu, M.: Chromatic characteristics of tongue color in tongue diagnosis of traditional Chinese medicine. *J. Guangzhou Univ. Tradit. Chin. Med.* **7**(4), 323–325 (2005)
14. Xu, Z.W., Chen, Q., Zhang, S.H.: A new research on chromaticity quantitative characteristics of cyan and purple tongue color. *J. Tradit. Chin. Med.* **22**(8), 1374–1375 (2004)
15. Weng, W., Huang, S.M.: Objective study on tongue diagnosis of traditional Chinese medicine. *Eng. Sci. China* **3**(1), 79–81 (2001)
16. Xu, Z.P., Zhang, B.L., Yao, Q.: Study on chromaticity of tongue in patients with breast cancer. *Sci. Technol. Chin. Med.* **7**(2), 67 (2000)
17. Chen, Y.Q., Xie, L.P.: Study on the relationship between tongue appearance and activity index of rheumatoid arthritis with dampness heat obstruction syndrome. *Liaoning J. Tradit. Chin. Med.* **40**(7), 1068–1070 (2013)

18. Fu, X.Y.: Quantitative analysis of tongue color in patients with rheumatoid arthritis and its clinical application [R]. Beijing University of Chinese Medicine (2018)
19. Yang, X.Y., Liang, R., Wang, Z.P., et al.: Research status and analysis of tongue color classification based on chromatics. *J. Beijing Univ. Tradit. Chin. Med.* **35**(8), 539–542,577 (2012)
20. Kawanabe, T., Kamarudin, N.D., Ooi, C.Y., et al.: Quantification of tongue colour using machine learning in Kampo medicine. *Eur. J. Integr. Med.* **8**(6), 932–941 (2016)
21. Li, Q., Liu, Z.: Tongue color analysis and discrimination based on hyperspectral images. *Comput. Med. Imaging Graph.* **33**(3), 217–221 (2009)
22. Wang, X.Z., Zhang, B., Yang, Z.M., et al.: Statistical analysis of tongue images for feature extraction and diagnostics. *IEEE Trans. Image Process.* **22**(12), 5336–5347 (2013)
23. Kamarudin, N.D., Ooi, C.Y., Kawanabe, T., et al.: A fast SVM-based tongue's color classification aided by k means clustering identifiers and color attributes as computer-assisted tool for tongue diagnosis. *J. Healthc. Eng.* <https://doi.org/10.1155/2017/7460168> (2017)
24. Tang, Y.P., Wang, L.R., He, X., et al.: Tongue image classification based on multi task convolution neural network. *J. Comput. Sci.* **45**(12), 255–261 (2018)
25. Jiao, W., Hu, X.J., Tu, L.P., et al.: Tongue color clustering and visual application based on 2D information. *Int. J. Comput. Assist. Radiol. Surg.* **15**(2), 203–212 (2020)
26. Zhang, Y.F., Hu, G.Q., Zhang, X.F.: Research on tongue color recognition algorithm of acne patients based on support vector machine. *J. Beijing Biomed. Eng.* **35**(1), 7–11 (2016)
27. Kaiming, H., Xiangyu, Z., Shaoqing, R., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
28. Sara, S., Nicholas, F.T., Geoffrey, E.H.: Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3859–3869. Neural Information Processing System Foundation, San Diego (2017)

# Investigation of Multi-task Learning for Object Detection



Yujie Zhang , Dongsheng Li , and Junping Xiang

**Abstract** Deep learning-based object detection is one of the most popular topics in computer vision. The generalization issue is particularly challenging due to the limited amount of labeled data and the difficulty in merging datasets in various domains. Multi-task learning is a promising way to leverage different datasets to enhance the generalization ability of networks, but it is rarely applied in object detection. In this paper, we focus on the effect of multi-task learning for one-stage object detection. Results show that the performance of the model trained in a small dataset can be improved significantly if it is trained jointly with a large dataset. The generalization of models trained with labeled data in a single domain can also be improved when trained jointly with different other domains.

**Keywords** Object detection · Feature enhancement · Multi-task learning

## 1 Introduction

Object detection is one of the most popular and challenging areas in computer vision and is under extensive investigation in academic research and practical applications. In object detection or even in the deep learning area, not only powerful algorithms are important, but also the abundant annotated data. Along with the thriving development of object detection algorithms, more and more datasets are being released. The availability of challenging and diverse object detection datasets provides tremendous support for the development of deep learning based object detection methods. In the past few years, some well-known object datasets have been released, such as PASCAL VOC [1], COCO [2], Objects365 [3], etc. The scale and variety of datasets also increase significantly.

---

Y. Zhang ()

Jiangsu Automation Research Institute of CSIC, Lianyungang, China  
e-mail: [zhangyujie@jari.cn](mailto:zhangyujie@jari.cn)

Y. Zhang · D. Li · J. Xiang

Lianyungang JARI Electronics Co., Ltd. of CSIC, Lianyungang, China

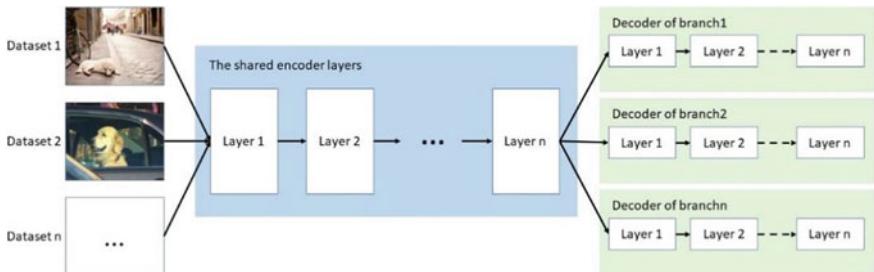
However, due to the variation of data collection methods, each dataset has its own unique features both in style and content [4]. These discrepancies lead to a distribution mismatch between one dataset and another. Therefore, every dataset can be regarded as laying in a unique domain. Existing detectors are usually domain-specific. One detector which performs well on the domain it is derived from may suffer a performance drop on samples in other domains [5], let alone real-world scenarios. This domain mismatch problem leads to the poor generalization ability of the object detector and is a big barrier to apply deep learning based detector to practice.

Multi-task learning is a training paradigm that could jointly learn different learning tasks simultaneously regardless of the type of different learning tasks and annotation format of datasets. In general, all or at least part of tasks are assumed to be related to each other [6, 7]. By sharing information about different related datasets, multi-task learning may help the neural network to learn domain-invariant representations, thus the network's generalization ability can be improved. However, rare investigations for one-stage object detection methods have been conducted or reported. Therefore, in this paper, we mainly focus on this issue and study the effect of multi-task learning in one-stage object detection learning.

## 2 Multi-task Learning Object Detector

The architecture of the neural network used in this paper is shown in Fig. 1. In this framework, the network consists of a common-shared encoder and several particular decoders. For convenience, we regard the shared encoder and a certain decoder as a branch. Thus, the encoder belongs to all of the branches in the network. Each of the branches is assigned to learn a certain task on a certain dataset.

The encoder can be any structure, such as VGG, ResNet, Darknet and so on. The encoder is mainly used for learning features from different datasets. By learning information on different datasets, the encoder is assumed to gain the potential to extract versatile features.



**Fig. 1** The architecture of our multi-task learning object detection network

The decoders in the neural network have almost the same structure, except for the number of filters of layers corresponding to output layers. Because the number of filters of prediction-related layers is related to the number of annotated categories, and different datasets may have different numbers of annotated categories. The decoders are responsible for mapping the features provided by the encoder to the final predictions.

### 3 Experiments

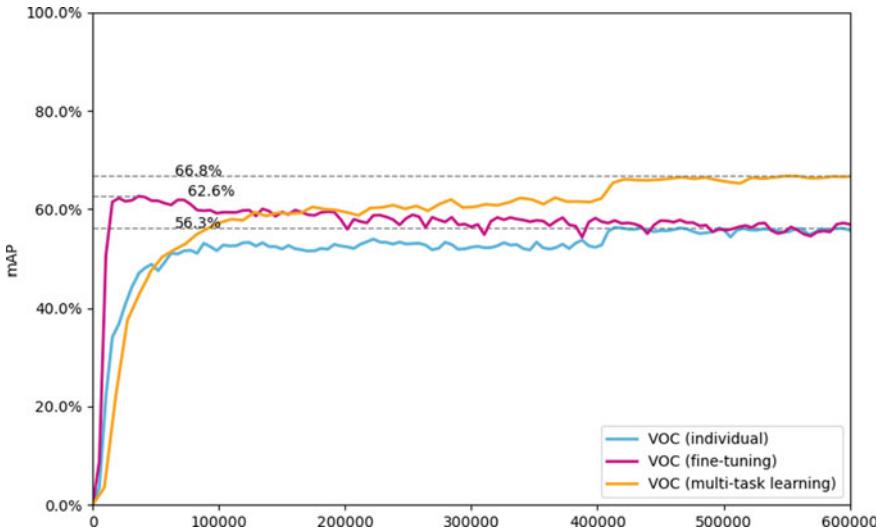
In this section, the performance of the multi-task learning approach on object detection was evaluated by examining two effects, namely the catastrophic issue and the generation improvement.

For the choice of the object detector, we select Yolov3-tiny as the baseline. To conduct the multi-task learning model, we add additional yolo layers for different branches. During the training process, one branch is randomly selected to be trained in one iteration, thus all branches are trained almost equal times in the whole training process.

#### 3.1 *Catastrophic Issue of Fine-Tuning Approach*

To evaluate the catastrophic issue, COCO is selected as the source domain, which is much larger and has more categories and VOC is selected as the target domain. For fine-tuning approach, we first train the detector on COCO dataset and then perform fine-tuning training on VOC dataset. For multi-task learning, a due multi-task learning network is built to train the model on COCO and VOC simultaneously. For comparison, we train one detector on VOC individually from scratch. The iteration-mAP curves of these networks on VOC are plotted in Fig. 2.

As shown in Fig. 2, the mAP of detector training individually from scratch achieves 56% of the mAP. The mAP of the fine-tuning method shows a sharp rise at the beginning of the training process, but it gradually falls backward and finally ends up with 57%, which is almost equal to the performance of detector training individually from scratch. The degradation of the performance of the fine-tuning method is caused by knowledge forgetting. Even though both training from scratch, multi-task learning obtains 67% of the mAP and achieves about 10% improvement. It indicates that learning with COCO jointly is helpful for improving the performance of the detector training on VOC.



**Fig. 2** Iteration-mAP curves of individual training, fine-tuning and multi-task learning

### 3.2 Generalization Improvement of Multi-task Learning Approach

To evaluate the generalization improvement of multi-task learning approach, three different traffic scene datasets are considered: UA-DETRAC (UA) [8], BDD100K (BDD) [9], and MIO-TCD (MIO) [10].

In this experiment, a multi-task learning network with three branches is designed to learn in these three datasets. For comparison, three other detectors are trained on UA, BDD, and MIO, individually, under the same condition and they are regarded as baseline models. After training, every branch of the multi-task learning network is validated on all three datasets to evaluate its generalization ability. The mAPs of all detectors in these three datasets are list in Table 1.

**Table 1** A comparison of mAP detectors obtained by baseline model and multi-task learning model on UA, BDD and MIO datasets

	Datasets		
	UA (%)	BDD (%)	MIO (%)
Baseline (UA)	72.19	4.34	32.49
Ours (UA)	<b>75.13</b>	<b>41.40</b>	<b>84.62</b>
Baseline (BDD)	30.24	<b>48.99</b>	16.16
Ours (BDD)	<b>58.51</b>	45.48	<b>76.06</b>
Baseline (MIO)	65.68	8.42	<b>93.91</b>
Ours (MIO)	<b>72.95</b>	<b>44.33</b>	90.29

For models trained on the UA dataset, the model trained individually (baseline) achieves high mAP (72.19%) on the UA dataset. But when it is tested on the other two datasets, it shows sharp mAP drops, with 4.34% mAP on BDD, and 32.49% mAP on MIO. The model trained by MTL achieves much higher mAP on all three datasets (75.13% on UA, 41.40% on BDD, and 84.62% on MIO). This suggests that the features learned by the neural network are augmented with multi-task learning approach and thus the generalization ability is enhanced. A similar situation can be observed on both BDD and MIO, except that there is a slight mAP drop for multi-task learning approach when evaluated on the same dataset of training individually.

## 4 Conclusion

In this paper, we focus on the effect of multi-task learning for object detection. The results show that by sharing parameters and features, the multi-task learning model training on a small scale dataset could obtain better performance when jointly learning on a dataset with a larger scale. Meanwhile, the multi-task learning method could enhance the generalization ability of detectors and provide a relatively more robust prediction when trained jointly with different datasets.

## References

1. Everingham, M., Gool, L.V., Williams, C., et al.: The Pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)
2. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer, Cham (2014)
3. Shao, S., Li, Z., Zhang, T., et al.: Objects365: a large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8430–8439 (2019)
4. Wilber, M.J., Fang, C., Jin, H., et al.: BAM! The behance artistic media dataset for recognition beyond photography. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1202–1211 (2017)
5. Chen, Y., Li, W., Sakaridis, C., et al.: Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3339–3348 (2018)
6. Thung, K.H., Wee, C.Y.: A brief review on multi-task learning. *Multimedia Tools Appl.* **77**(2), 29705–29725 (2018)
7. Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint [arXiv:1707.08114](https://arxiv.org/abs/1707.08114) (2017)
8. Wen, L., Du, D., Cai, Z., et al.: UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.* **193**, 102907 (2020)
9. Yu, F., Chen, H., Wang, X., et al.: BDD100K: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 (2020)
10. Luo, Z., Frederic, B.C., Carl, L., et al.: MIO-TCD: a new benchmark dataset for vehicle classification and localization. *IEEE Trans. Image Process.* **27**(10), 5129–5141 (2018)

# Design of Place Recognition Algorithm Based on VLAD Code and Convolutional Neural Network



Bo Wang , Xinsheng Wu , An Chen , and Hongxia Gao

**Abstract** Visual place recognition is an important and challenging research topic. With the increasing complexity of vision recognition tasks, the traditional image recognition algorithm can not deal with the problem of large-scale illumination change and the interference caused by object occlusion in image. In recent years, deep learning has made many achievements. The main solution proposed based on the deep learning method is to design an end-to-end recognition network, and use pre-labeled place datasets to train the network and finally obtain the classification results of the network. In this paper, a network of place recognition algorithm is proposed which combines deep convolutional features. The sensing field is expanded by adding convolution module, and the extracted convolution features are sent to the local aggregation vector module to obtain the place description vector. The training result of Tokyo Time Machine dataset and Pittsburgh dataset shows that the improved algorithm proposed in this paper has a better recognition recall rate. The generalization performance of the general place retrieval dataset, such as Oxford architecture and Paris architecture, is better than that of the contrast algorithm.

**Keywords** Deep learning · Visual place recognition · Image descriptor

## 1 Introduction

Images are an important form of information expression. In recent years, vision-based autonomous driving technology has developed rapidly, and vision-based real-time positioning and map construction algorithms VSLAM (Visual Simultaneous Localization and Mapping) have emerged. The algorithm includes four parts: visual odometry, back-end map optimization, closed-loop detection, and map construction.

---

B. Wang · X. Wu

School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China

A. Chen · H. Gao

South China University of Technology, Guangzhou 510640, China

e-mail: [chenan@scut.edu.cn](mailto:chenan@scut.edu.cn)

The function of closed-loop detection is to judge whether the image obtained by the current visual odometer is a known node. The accurate closed-loop judgment result can add constraint conditions for the back-end optimization, optimize the motion mode of the robot, and eliminate the accumulated error of the visual odometer [1]. The failure of loopback detection will affect the construction of the global map and ultimately affect the navigation effect. Therefore, designing a more effective place recognition algorithm and improving the speed and accuracy of recognition is of great significance to the loop detection link of the SLAM algorithm.

After the research of scholars, the place recognition algorithm has achieved some results. However, the proposed methods all face the problem of low recognition accuracy. The reason is that the diversity of foreground objects in the real scene and the complexity of the spatial structure of the background environment make the pictures of the same place have great differences [2]. At the same time, changes in ambient light intensity and viewing angle during shooting will also have a certain impact on the recognition algorithm. How to improve the recognition effect of the place recognition algorithm in complex situations is the focus of this article.

## 2 Related Methods

The steps of traditional place recognition algorithms are image preprocessing, image feature extraction, feature descriptor construction, feature dimensionality reduction and classifier design [2]. The key step is image feature extraction. Traditional methods use hand-designed feature descriptions in the process of extracting image features. The main recognition algorithms include Bag of Visual Words (BOV), Fisher Vector (FV), and Vector of Locally Aggregated Descriptor (VLAD).

The BOV algorithm extracts Scale Invariant Features Transform (SIFT) features for each image patch. Then cluster the features of the image. The image to be tested performs frequency statistics on the features according to the word list formed by the cluster centers [3]. Fisher vector algorithm uses the gradient vector of the likelihood function to represent the characteristics of an image. The algorithm first calculates the SIFT features of the image, and then calculates the sum of the normalized gradient statistics of the extracted T SIFT features that obey the same distribution as a descriptor [4]. The VLAD algorithm first calculates the features of the image and obtains the cluster center of the image through the Kmeans algorithm [5]. Then calculate the sum of the residuals of each dimension from the feature point of the image to the cluster center, and obtain the image descriptor with the difference between the feature point and the cluster center as the word list [1]. Compare and select the nearest category result as the category of the image to be tested [6].

In recent years, deep learning has achieved great achievement in many fields. The development of deep learning has promoted the improvement of place recognition methods. The use of convolutional neural networks for place recognition has become a current trend. Arandjelovic et al. proposed the NetVLAD algorithm, which is mainly to design an end-to-end recognition network. The positive samples

and negative samples are input into the classification network, and the optimal place recognition results are obtained by learning the convolution kernel parameters. Through training on the Tokyo Time Machine dataset, an optimal recognition effect is obtained. Hou uses the results obtained on the general scene recognition network PlaceCNN for place recognition, which has a higher recognition effect in scenes with changing lighting [7]. But the algorithm takes a long time, and at the same time, the recognition effect of the place picture of the driving environment is average.

### 3 Design of Place Recognition Algorithm

#### 3.1 Trainable VLAD Coding Model

Getting ideas from the VLAD method and NetVLAD algorithm in traditional place recognition methods, this paper proposes an improved Netvlad algorithm combines deep convolution features. By extracting high-level semantic features of location pictures, the accuracy of place recognition is higher. For the VLAD algorithm, assume that the image is calculated to obtain N D-dimensional feature vectors  $x_i$ . The number of clusters is K, Center is  $c_k$ . Then the algorithm outputs a  $K \times D$  dimensional image feature vector. The vector at the center of the cluster  $a_k$  is 1, the parameter not in the case of cluster center is 0. The VLAD vector is calculated as follows:

$$V(j.k) = \sum_{i=1}^N a_k(x_i)(x_i(j) - c_k(j)), \quad k \in K, j \in D \quad (1)$$

According to the proposal: and the continuous value model is given when the  $a_k$  parameter in the VLAD algorithm is a discrete value. Given parameter  $a_k$  to multiple cluster centers. The parameter value corresponding to the feature closer to the cluster center is closer to 1, and the parameter value corresponding to the feature farther from the cluster center is closer to 0 [8]. So as to satisfy the principle of differentiable neural network parameters. The VLAD coefficient is calculated as follows:

$$\bar{a}(x_i) = \frac{e^{-\alpha \|x_i - c_k\|^2}}{\sum_{k'} e^{-\alpha \|x_i - c_{k'}\|^2}} \quad (2)$$

Expanding the square term in Formula (2), and subtracting the exponential term from the numerator and denominator at the same time, the coefficient distribution formula is as follows:

$$\bar{a}_k(x_i) = \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'}^T x_i + b_{k'}}} \quad (3)$$

where: vector  $w_k = 2\alpha c_k$ ,  $b_k = -\alpha \|c_k\|^2$ . Incorporating the above formula into Formula (1), the image feature expression relationship extracted by the deep neural network is shown as follows:

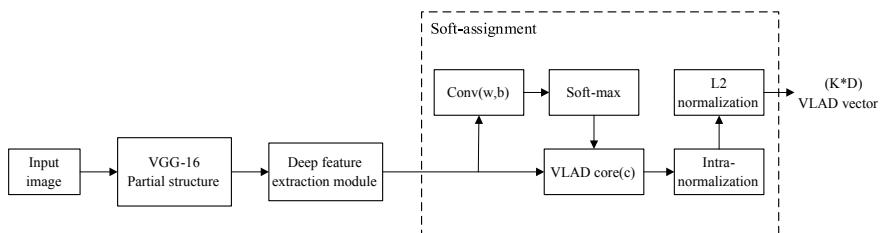
$$V(j, k) = \sum_{i=1}^N \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'}^T x_i + b_{k'}}} (x_i(j) - c_k(j)) \quad (4)$$

where  $w_k$ ,  $b_k$ ,  $c_k$  are parameters that can be learned through the network. Using stochastic gradient descent algorithm for learning can obtain the optimal parameters of the network.

### 3.2 NetVLAD Feature Extraction Network Fused with High-level Features

In the problem of place recognition, the images of the same location usually show great differences due to the movement of dynamic objects or the change of perspective. In order to make the features more suitable for the problem of place recognition, this paper expands the receptive field of the input image based on the existing VGG16 feature network extraction, and adds a feature extraction module, so that the deep learning network can extract the deep image Semantic information. Connect the VLAD module after the semantic feature extraction module to calculate local feature aggregation vector. The extracted feature vector is given by the normalized feature.

The specific network model is shown in Fig. 1. On the basis of the third convolution block of the VGG-16 network of the original NetVLAD model, a feature extraction module is added. The size is the same as that of the third convolution block, which is  $56 \times 56$ , and the channel is also 256, using to extract the features of the third convolution block. And send the output result to the fourth layer of convolution block. The fourth layer convolution module adds a convolution layer to extract the features of this layer, the size is the same as conv4,  $28 \times 28 \times 512$ , and the output result is sent to the fifth layer convolution layer. Add a feature extraction module on



**Fig. 1** Improved place recognition network flow chart

the basis of the original VGG16 feature extraction on the fifth convolutional layer, the size is the same as that of conv5, which is  $14 \times 14 \times 512$ . The feature matrix calculated by the deep semantic extraction module is sent to the NetVLAD layer for classification.

The soft-assignment module can be divided into 4 parts. First, the conv layer performs convolution calculation on the output of the semantic feature extraction module, and the obtained result passes through the soft-max layer to obtain the coefficient matrix of netvlad. The output of the high-level semantic feature extraction layer in the other channel is directly used as the input of the VLAD kernel to calculate the residuals of the feature distance from the cluster center. Multiply the results of them to obtain a feature vector representing, and obtain a vector of size  $K \times D$  as a place representation through normalization.

The improved recognition network adds three convolution modules for extracting deep features. The receptive field is enlarged, The processing flow of our method is shown as follows:

For the depth feature extraction module added by the third convolution block, Given symbol  $a_q^7$  is the forward output of the 7th layer after the activation function.  $k_{p,q}^8$  is the convolution kernel corresponding to the  $p$  channel of the 8th layer and the  $q$  channel of the 7th layer.  $b_p^8$  is the bias of the  $p$ -channel on the 8th layer. The extracted feature map is shown as follows:

$$z_p^8(i, j) = \sum_{q=1}^{256} \sum_{u=-1}^1 \sum_{v=-1}^1 a_q^7(i - u, j - v) k_{p,q}^8(u, v) + b_p^8 \quad (5)$$

For the deep feature extraction module added by the fourth convolution block, Given symbol  $a_q^{11}$  is the forward output of the 11th layer after the activation function.  $k_{p,q}^{12}$  is the convolution kernel corresponding to the  $p$  channel of the 12th layer and the  $q$  channel of the 11th layer,  $b_p^{12}$  is the bias of the  $p$ -channel on the 12th layer. The extracted feature map is shown as follows:

$$z_p^{12}(i, j) = \sum_{q=1}^{512} \sum_{u=-1}^1 \sum_{v=-1}^1 a_q^{11}(i - u, j - v) k_{p,q}^{12}(u, v) + b_p^{12} \quad (6)$$

For the depth feature extraction module added by the fifth convolution block, Given symbol  $a_q^{15}$  is the forward output of the 15th layer after the activation function.  $k_{p,q}^{16}$  is the convolution kernel corresponding to the 16th layer  $p$  channel and the 15th layer  $q$  channel.  $b_p^{16}$  is the bias of the  $p$ -channel on the 16th layer. the extracted feature map is shown as follows:

$$z_p^{16}(i, j) = \sum_{q=1}^{512} \sum_{u=-1}^1 \sum_{v=-1}^1 a_q^{15}(i - u, j - v) k_{p,q}^{16}(u, v) + b_p^{16} \quad (7)$$

**Table 1** Comparison of the receptive field of VGG16 and this algorithm

VGG16 (Unit: pixel)	Our algorithm (Unit: pixel)
196	<b>252</b>

The bold data were determined experimentally and quoted from the literature 9

Our method can extract higher-dimensional image features. Given symbol  $RF_i$  is the receptive field of the  $i$ -th convolutional layer.  $RF_{i+1}$  is the receptive field of  $i + 1$  layer, stride is the step size of the convolution, and  $K_{size_i}$  is the size of the convolution kernel of the current layer. The calculation of the receptive field is shown as follows:

$$RF_i = (RF_{i+1} - 1) \times stride_i + K_{size_i} \quad (8)$$

The comparison of the receptive field size of the original paper network receptive field and the improved network proposed in this paper is shown as following (Table 1):

### 3.3 Training of NetVLAD Network Fused with High-level Features

The Tokyo Time Machine is a dataset taken by Google Maps. Each position in the query set of this dataset is an image captured during the day, sunset and night of the day [8]. Different from the image classification dataset, the images taken by the Google Street View dataset are full-view images. Each panoramic image is composed of a set of 12 perspective images in different directions, and these images are uniformly taken in different directions [8]. Each image only contains GPS tags that represent its approximate location on the map. The dataset contains 49,104 training pictures, 49,056 verification pictures and 75,984 test images.

The Pittsburgh dataset is a dataset captured in the city of Pittsburgh. The feature of the dataset is repeated external structures, such as building exterior walls, road fences and road marking information. The training set contains 7416 pictures, the verification set contains 7608 pictures, and the test set contains 6816 pictures.

A supervised triple ordering loss function is used to deal with incomplete data and noisy location annotations in street scenes. For a given image  $q$  to be queried, it is necessary to find an image  $I_i^*$  with a higher ranking than other images in the training set. Set the image whose GPS distance in the dataset is within 25 m from the current image as a positive sample set  $\{p_i^q\}$ . Images with a distance of more than 25 m are used as a set of negative samples  $\{n_j^q\}$ . It is converted into a ternary array sort loss  $\{q, p_i^q, n_j^q\}$ . For the image to be trained, find the image  $p_{i^*}^q$  with the shortest distance in the set of positive samples. We can get the following Formula (9):

$$p_{i^*}^q = \arg \min_{p_i^q} d_\theta(q, p_i^q) \quad (9)$$

It is required that the distance  $d_\theta$  to the positive query dataset is less than the distance to all negative query datasets, We can get the following Formula (10):

$$d_\theta(q, p_{i^*}^q) < d_\theta(q, n_j^q), \quad \forall j. \quad (10)$$

According to the above relationship, the weakly supervised ranking loss function is designed [8]. Given parameter  $l$  is the Hingel loss function used to train the classifier.  $m$  is an additional constant.  $L_\theta$  represents the sum of the loss of all negative sample images. When the distance between the negative sample image and the query image is greater than the sum of the distance between the query image and the expected best matching image and the additional constant  $m$ , the negative distance loss is set to 0, When the positive distance is greater than the threshold, the loss needs to be increased proportionally. The loss function is shown as follows:

$$L_\theta = \sum_j l((\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_j^q))) \quad (11)$$

In the process of backpropagation, stochastic gradient descent (SGD) is used to train the parameter  $\theta$  expressed by the network. Set the number of samples in each training batch to 2 triples. Momentum is 0.9. The initial learning rate is 0.0001. The learning rate is halved every 5 times of iterative learning to prevent oscillations at the extreme points. Training time is 30. Use the trained model to verify on the Tokyo Time Machine validation set. The result shows that when N is equal to 10, Recall@N is equals to 0.98. Better than the current best place recognition model.

## 4 Experiment

In order to verify the performance of the improved method proposed in this article, it is compared with other closed-loop detection algorithms that apply place recognition algorithms. The environment configuration on Tokyo streetscape dataset is: 32 GB memory, a piece of GeForce RTX 2080 SUPER graphics card, the processor is Intel Core i5-9400F, the operating environment is win10 and Matlab 2016. The environment configuration on the Pittsburgh dataset is: 2 Graphics cards Geforce RTX 2080 Ti which memory is 11G, CPU is Intel(R) Xeon(R) Gold 5218. The framework is pytorch and CUDA10.2

The general place retrieval datasets used include Oxford Buildings Datasets and Paris Buildings Datasets and INRIA Holidays. Where Oxford Buildings Datasets Contains 5062 images obtained by searching for specific Oxford landmarks. Generate comprehensive ground truth for 11 different landmarks. The Paris dataset contains

6412 images obtained by searching for specific Paris landmarks, including 12 categories. The INRIA Holidays dataset has a total of 1491 images. Contains a wealth of posture rotation, viewpoint and illumination changes and other interference images.

The evaluation indicators are indicators commonly used in place recognition algorithms. If at least one of the first N database images retrieved is 25 m away from the real position of the query, it is considered that the query image has been correctly positioned. Then draw the Recall value of the correct recognition result for different values. The recall rate is calculated as follows:

$$R = \frac{TP}{TP + FN} \quad (12)$$

TP represents the number of positive classes predicted as positive classes, and FN represents the number of negative classes predicted as positive classes. In order to verify the accuracy of the algorithm, the accuracy of different algorithms for retrieving datasets in general locations was tested. FP means the number of negative samples predicted to be positive. The accuracy rate is calculated as following:

$$P = \frac{TP}{TP + FP} \quad (13)$$

The comparison of the test result of the algorithm trained on Tokyo Time Machine dataset is shown in Table 2, where the off-the-shelf algorithm is the result of using the pretrain model of the NetVLAD algorithm, and NetVLAD is the experimental algorithm for the algorithm comparison mentioned in this paper.

Through the following comparison, it can be found that the improved NetVLAD algorithm proposed in this paper has the same overall effect as NetVLAD, and it is higher than the NetVLAD algorithm when recall@N is equal to 10. It can be seen from Figs. 2 and 3 that when N is greater than 10, it is also higher than the NetVLAD algorithm. It shows that this algorithm has better results when the value of N is larger.

The accuracy of different recognition algorithms trained on Tokyo Time Machine and test on the general test set is shown in Table 3. It can be seen from the table that the recognition accuracy of deep learning methods is mostly higher than that of traditional methods. In addition, the improved NetVLAD algorithm fused with deep semantic features in this experiment has a higher accuracy than the NetVLAD algorithm and comparison algorithm.

**Table 2** Statistics of recall rates of different algorithms

Algorithm	recall@1	recall@10
Off-the-shelf	0.91	0.97
NetVLAD	0.95	0.97
Our algorithm	<b>0.95</b>	<b>0.98</b>

The bold data were determined experimentally and quoted from the literature 9

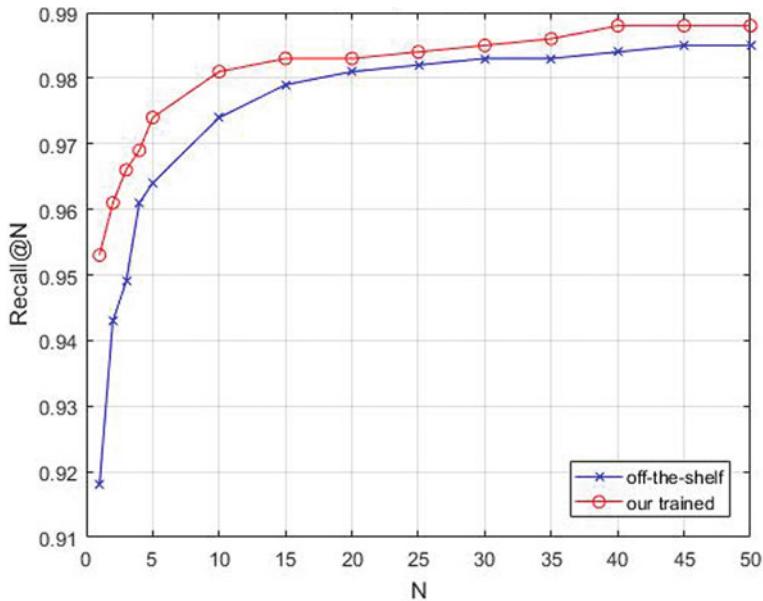


Fig. 2 Recall@N-N image of our method

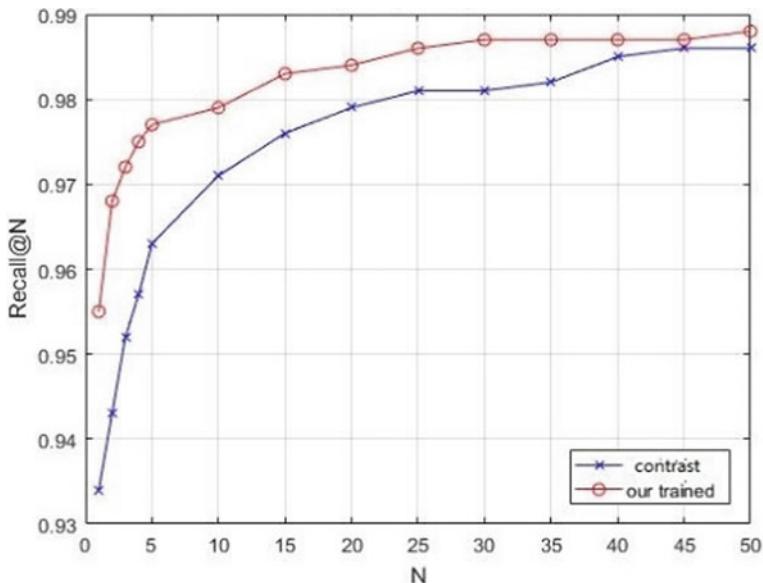


Fig. 3 Original algorithm Recall@N-N image

**Table 3** Accuracy statistics of different test sets

Dataset	Improved BOV	NetVLAD Off-the-shelf	NetVLAD	Our algorithm
Oxford buildings	52.3	55.5	63.5	<b>72.45</b>
Paris buildings	70.1	67.7	77.9	<b>78.28</b>
INRIA holidays	–	80.1	79.9	<b>82.59</b>

The bold data were determined experimentally and quoted from the literature 9

The comparison of the test result of the algorithm trained on Pittsburgh dataset is shown in Table 4. The comparison algorithm is based on the model reproduced by NetVLAD algorithm, and spe-vlad algorithm is the algorithm proposed by Yu [9]. The comparison experiment results show that the recall rate of the algorithm proposed in this paper is higher than that of the comparison algorithm.

The accuracy of different recognition algorithms trained on Pittsburgh dataset and test on the general test set is shown in Table 5. The easy model of Oxford building dataset and Paris building dataset is used to carry out the experiment [10]. From Table 5, it can be seen that the model trained in Pittsburgh dataset by this method has a 3 percentage point higher accuracy in Oxford building dataset than netvlad. It is basically the same as the original method in Paris building dataset. Table 6 shows the recall rate statistics in general place retrieval datasets. Shown that our model has better recall performance.

**Table 4** Statistics of recall rates of different algorithms

Algorithm	recall@1	recall@10	recall@20
SPE-VLAD	0.79	0.91	0.94
NetVLAD	0.8222	0.9376	0.9553
Our algorithm	<b>0.8327</b>	<b>0.9442</b>	<b>0.9599</b>

The bold data were determined experimentally and quoted from the literature 9

**Table 5** Accuracy statistics of different test sets

Dataset	NetVLAD	Our algorithm
Oxford buildings	55.66	<b>58.66</b>
Paris buildings	<b>74.22</b>	74.03

The bold data were determined experimentally and quoted from the literature 9

**Table 6** Accuracy statistics of different test sets

Dataset	NetVLAD (%)	Our algorithm (%)
Oxford buildings	86.76	<b>89.71</b>
Paris buildings	98.57	<b>98.57</b>

The bold data were determined experimentally and quoted from the literature 9

**Table 7** Comparison of running time of different algorithms

Algorithm	Dimension	Feature extraction time/s
BOW	1024	—
NetVLAD	4096	0.03
Our algorithm	4096	0.11

The algorithm time efficiency comparison is shown in Table 7. Because BOW uses hand-designed descriptors, the time consumption of extracting descriptors in the same experimental environment as the deep learning algorithm is negligible, and the other two methods meet the approximate real-time performance. The improved method in this paper takes slightly higher average time for feature extraction in the test set than the original algorithm.

## 5 Conclusion

This paper proposes a place recognition algorithm fused with deep semantic features. Added convolution operation modules on the basis of the original deep learning NetVLAD network, expanded the receptive field and extracted the deep features of the image. Using Tokyo Time Machine and Pittsburgh dataset for training. And using the ternary array ranking loss as the loss function, the optimal parameters are obtained by learning through the stochastic gradient descent method. The optimal network model obtained by training is tested on general place retrieval datasets such as Oxford Building and Paris Building datasets. The experimental results show that the place recognition method proposed in this paper improves the accuracy of recognition basically satisfies the real-time nature of closed-loop detection. The next step will focus on how to integrate the results of the image segmentation algorithm to make the algorithm more robust to large-scale changes in illumination and viewing angles.

**Acknowledgements** This paper is supported by the Natural Science Foundation of Guangdong Province, China (Grant No. 2019A1515011041), National Natural Science Foundation (61873096, 62073145), Guangdong Province Basic and Applied Basic Research Fund Project (2020A1515011057), Guangdong International Cooperation Fund Project (2020A0505100024), Central University Project (D2201200) and Xijiang Innovation Team Project.

## References

1. Ang, L., Xiaogang, R., Jing, H., Xiaoqing, Z.: Closed-loop detection method combining convolutional neural network and VLAD. Comput. Appl. Softw. **38**(01), 135–142 (2021)
2. Ran, M., Feifei, L., Qiu, C.: Scene recognition method based on convolutional neural network and multi-scale spatial coding. Electron. Sci. Technol. **33**(12), 54–58,74 (2020)

3. Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. Cybern.* **1**(1–4), 43–52 (2010)
4. Sánchez, J., Perronnin, F., Mensink, T., et al.: Image classification with the fisher vector: theory and practice. *Int. J. Comput. Vis.* **105**(3), 222–245 (2013)
5. Picard, D., Gosselin, P.H.: Improving image similarity with vectors of locally aggregated tensors. In: 2011 18th IEEE International Conference on Image Processing. IEEE, pp. 669–672 (2011)
6. Zhao, Y.: Research on Visual Feature Expression and Learning for Image Classification and Recognition. South China University of Technology (2014)
7. Hou, Y., Zhang, H., Zhou, S.: Convolutional neural network-based image representation for visual loop closure detection. In: 2015 IEEE International Conference on Information and Automation. IEEE, pp. 2238–2245 (2015)
8. Arandjelovic, R., Gronat, P., Torii, A., et al.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297–5307 (2016)
9. Yu, J., Zhu, C., Zhang, J., et al.: Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(2), 661–674 (2019)
10. Radenović, F., Iscen, A., Tolias, G., et al.: Revisiting Oxford and Paris: large-scale image retrieval benchmarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5706–5715 (2018)

# Fusing Global Gabor Feature and Local Binary Pattern for Texture Image Recognition



Junmin Wang 

**Abstract** Texture is a fundamental visual feature in human vision, and texture image recognition is an important part of artificial intelligence. However, traditional texture recognition methods can't effectively utilize global and local texture features simultaneously, which restricts the recognition performance of traditional methods. Therefore, this work combines global Gabor feature and local binary pattern to further improve the performance of texture image recognition. Firstly, a multi-scale image pyramid space is generated to reflect the scale variation of texture image. Secondly, Gabor filtering is used in the image pyramid space to extract the global Gabor feature, which is used as the global texture feature. Thirdly, the completed local binary count algorithm with multiple radii is applied to original image to extract the local binary pattern, which is used as the local texture feature. Finally, the extracted global and local texture features are fused to recognize the texture image by the nearest subspace classifier (NSC). The experimental results show that the proposed method can achieve state-of-the-art recognition accuracy with high efficiency. In addition, the proposed method is robust to scale variation and the number of training samples in texture recognition task.

**Keywords** Texture classification · Feature extraction · Completed local binary count · Gabor filtering

## 1 Introduction

Texture is ubiquitous in the real world, and the texture patterns contained in texture images play an important role in human's perception of the real world. Therefore, it is of great significance to do researches on the texture image recognition (classification). At present, texture image recognition has gained huge success in facial expression recognition [1], remote sensing image classification [2], medical image analysis [3], material classification [4] and so on. However, texture image recognition is very

---

J. Wang 

School of Information Engineering, Pingdingshan University, Pingdingshan 467000, China  
e-mail: [wjunmin2000@mail.nwp.edu.cn](mailto:wjunmin2000@mail.nwp.edu.cn)

difficult because the texture image in different scenarios may vary significantly due to object rotation, viewing and illumination changes, scale variations, etc.

A large variety of texture recognition methods have been proposed, and the main three categories include:

1. Local binary pattern (LBP) based methods. The LBP algorithm proposed by Ojala et al. [5] is very popular due to its theoretical simplicity, high recognition accuracy and computational efficiency. Guo et al. [6] proposed the completed LBP (CLBP) method by introducing the difference magnitude and the gray-level of center pixel, which significantly improved the recognition accuracy. Later on, Zhao et al. [7] proposed the completed local binary count (CLBC) by abandoning the local binary structural information to significantly reduce the computational cost. The scale selective LBP (SSLBP) [8] method proposed by Guo et al. adopted multi-scale space and dominant patterns to enhance the robustness to scale variation.
2. Texton learning based methods. Texton is the primitive feature unit of a texture image. Varma and Zisserman [9] proposed the VZ-MR8 method which used a filter bank to learn a set of textons. Later on, Varma and Zisserman [10] further proposed the VZ-Joint method, which directly used the image patches to learn the textons. Recently, Xie et al. [11] proposed the texton encoding induced statistical features (TEISF) method which used the l2-norm regularization to learn a texton dictionary.
3. Gabor filtering based methods. Gabor wavelets have good properties in imitating the function of simple cortical cells [12, 13], which can extract the texture features with different scales and orientations. Therefore, Manjunath and Ma [14], and Arivazhagan et al. [15] used a Gabor filter bank to extract texture information. The local Gabor wavelets binary patterns (LGWPB) method [16] used the Gabor filtering and LBP for texture description. More recently, Wang et al. proposed the GLGF method [17] where the joint encoding of magnitude and phase of Gabor filtered image was used as the texture feature for texture classification.

The main contributions of this study are as follows:

1. Higher recognition accuracy with high efficiency. The global Gabor feature and local binary pattern are extracted separately to represent the global and local texture feature respectively, and then the above two features are further fused by the NSC classifier, which achieves state-of-the-art texture recognition accuracy with high efficiency.
2. Scale robustness. A multi-scale image pyramid space is constructed to describe the images with different scales for global Gabor feature extraction, and multiple radii (scales) are selected for local binary pattern extraction, which can overcome the scale changes of texture images.

The paper is organized as follows. The traditional Gabor filtering and CLBC algorithm are introduced in Sect. 2. The principles of the proposed method are explained

in Sect. 3. The experimental results of the proposed method and related performance analysis are shown in Sect. 4. Some conclusions are given in Sect. 5.

## 2 Related Work

### 2.1 Gabor Filtering

A Gabor function  $\mathbf{g}(x, y)$  in the spatial domain can be defined as [14]:

$$\mathbf{g}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right] \exp(2\pi j Fx) \quad (1)$$

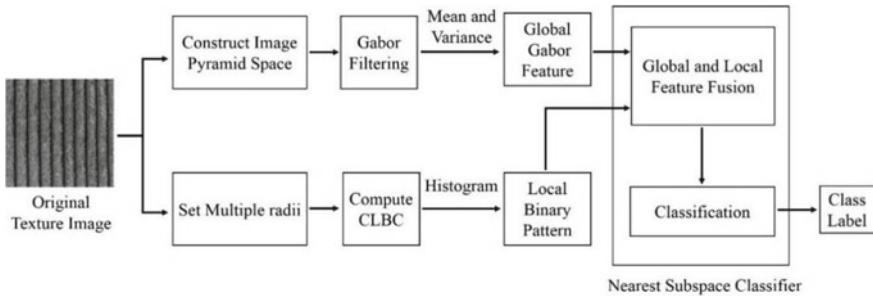
where  $F$  is the spatial frequency of interest,  $\sigma_x$  and  $\sigma_y$  represent the standard deviations in  $x$  and  $y$  directions. The desired Gabor filter bank is generated by implementing scaling and rotation operations to  $\mathbf{g}(x, y)$ . After that, the obtained Gabor filter bank can be used to extract the texture features corresponding to different scales and orientations as follows:

$$\mathbf{W}^{s,k}(x, y) = \mathbf{I}(x, y) * \mathbf{g}^{s,k}(x, y), \quad s = 1, 2, \dots, S, k = 1, 2, \dots, K \quad (2)$$

where  $\mathbf{I}(x, y)$  is the original image,  $\mathbf{g}^{s,k}(x, y)$  is the Gabor filter at scale  $s$  and orientation  $k$ ,  $*$  denotes the convolution operation,  $\mathbf{W}^{s,k}(x, y)$  is the Gabor filtered image,  $K$  and  $S$  denote the number of orientations and scales respectively. Finally, the mean and variance of magnitude of  $\mathbf{W}^{s,k}(x, y)$  at multiple orientations and scales are adopted to characterize the texture information for texture recognition.

### 2.2 CLBC Algorithm

Zhao et al. [7] found that the most discriminative information of local texture for rotation invariant texture classification was not the ‘micro-structures’ information but the local binary grayscale difference information, therefore they discarded the structural information from LBP operator and proposed the local binary count (LBC) operator. Meanwhile, motivated by the CLBP, they further proposed a completed LBC (CLBC). Compared with the CLBP, the CLBC can achieve comparable classification accuracy with lower computational complexity.



**Fig. 1** The scheme of the proposed method

### 3 Proposed Method

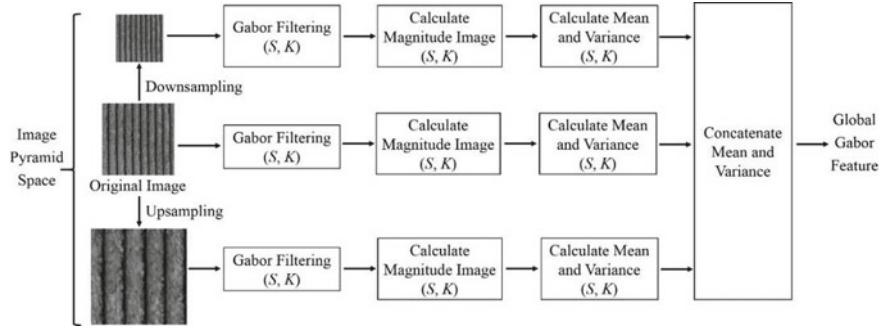
#### 3.1 The Proposed Scheme

The scheme of the proposed method is shown in Fig. 1. In the proposed method, the global Gabor feature and local binary pattern are extracted separately. On the one hand, to extract the global Gabor feature, a multi-scale image pyramid space is generated by sampling and interpolating the original texture image, and then all the images in the multi-scale pyramid space are convolved with the Gabor filters, and finally the mean and variance of filtered magnitude images are concatenated to describe the global Gabor feature. On the other hand, to extract the local binary pattern, the CLBC algorithm with multiple radii is applied to the original image, and then the joint histogram is used as the local texture feature. At last, the extracted global and local texture features are fused for the final texture image recognition by the NSC classifier [11].

#### 3.2 Global Gabor Feature Extraction

To obtain discriminative and scale-robust global texture features, we propose the following global Gabor feature extraction scheme, which is shown in Fig. 2.

1. Generating a multi-scale image pyramid space for each original texture image. When the distance between texture sample and camera varies, the acquired texture images exhibit significantly different visual characteristics, which greatly increases the difficulty of texture image recognition. To address the scale variation issue of texture recognition, we construct a pyramid space for each original image by sampling and interpolating the original image. Therefore, we obtain an image pyramid space with three levels, which can imitate the scale variation of texture images and the features derived from three images in



**Fig. 2** The scheme of global Gabor feature extraction

the pyramid space is more robust to scale variation than that derived from only the original image.

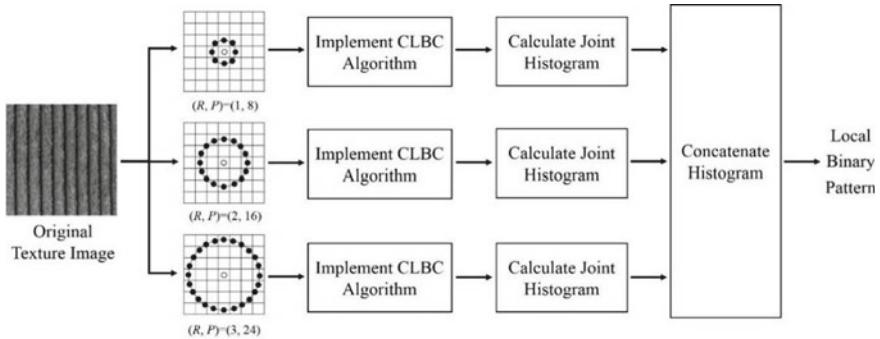
2. Implementing Gabor filtering in the image pyramid space. We adopt the method proposed by Manjunath and Ma [14] to design a bank of Gabor filters. Then, we implement the Gabor filtering by the convolution operation between the images in the image pyramid space and the designed Gabor filter bank. Obviously, the image pyramid space corresponding to an original texture image has three images, and the Gabor filter bank has  $S \times K$  Gabor filters. Therefore, for an original image, the total number of Gabor filtered images is  $3 \times S \times K$ .
3. Extracting Global Gabor feature. After Gabor filtering, the mean and standard deviation of magnitude of  $3 \times S \times K$  Gabor filtered images are calculated and concatenated as follows:

$$\mathbf{h}_{global} = \left[ \mu_1^{1,1}, \sigma_1^{1,1}, \mu_1^{1,2}, \sigma_1^{1,2}, \dots, \mu_1^{S,K}, \sigma_1^{S,K}, \dots, \mu_3^{1,1}, \sigma_3^{1,1}, \mu_3^{1,2}, \sigma_3^{1,2}, \dots, \mu_3^{S,K}, \sigma_3^{S,K} \right] \quad (3)$$

where  $\mu_i^{s,k}$  is the mean value,  $\sigma_i^{s,k}$  is the standard deviation value, the superscript  $s$  and  $k$  represent the corresponding scale and orientation, and the subscript  $i$  means the level of the pyramid space. The obtained  $\mathbf{h}_{global}$  is considered as the global feature of original texture image. Obviously, the proposed global Gabor feature extraction scheme generates a multi-scale image pyramid space, which captures more texture features with different scales and is more robust to the scale change of texture images.

### 3.3 Local Binary Pattern Extraction

To obtain discriminative and scale-robust local texture features, we propose the following local binary pattern extraction scheme, which is illustrated in Fig. 3.



**Fig. 3** The schematic diagram of local binary pattern extraction

The extraction of local binary pattern is based on the CLBC algorithm. First, we select three values for  $(R, P)$  parameter pair, namely  $(R, P)=(1, 8)$ ,  $(2, 16)$ , and  $(3, 24)$ , where the small radius is used to capture the texture feature with small size, and the big radius is used to capture the texture feature with big size. Second, the CLBC algorithm is implemented on the selected  $(R, P)$  parameter pair, and the joint histogram CLBC\_SMC is calculated, where CLBC\_SMC is the joint histogram of CLBC\_S, CLBC\_M and CLBC\_C components. Finally, the joint histogram CLBC\_SMC on three radii are concatenated as follows:

$$\mathbf{h}_{local} = [CLBC\_SMC_1^8, CLBC\_SMC_2^{16}, CLBC\_SMC_3^{24}] \quad (4)$$

which is considered as the local texture feature of original texture image.

### 3.4 Feature Fusion and Recognition

The global Gabor feature  $\mathbf{h}_{global}$  mainly describe the holistic texture feature, and the local binary pattern  $\mathbf{h}_{local}$  can characterize the locally detailed feature of texture image, which are both important for texture image representation and recognition. However, most methods only use either  $\mathbf{h}_{global}$  or  $\mathbf{h}_{local}$  for texture recognition, which restricts the performance of these methods. Therefore, to effectively utilize the global and local features and boost the performance of texture recognition, we propose to fuse the global Gabor feature and local binary pattern, and implement the final texture recognition in the framework of NSC [11].

## 4 Experiments and Results

### 4.1 Datasets

Two benchmark texture datasets, namely CUReT dataset [8] and KTH-TIPS dataset [8], are used to evaluate the proposed method. The CUReT dataset includes 61 texture classes and each class has 92 image samples. The KTH-TIPS dataset includes 10 classes of texture sample and each class includes 81 images, and there are a total number of 810 images.

### 4.2 Experimental Setting

All images were first converted to grayscale images before feature extraction, therefore the color information was not used to discriminate between different textures. In the following experiments, set the value of related parameter as follows: Gabor filter support size  $N_g = 15$ ,  $S = 4$ ,  $K = 6$ ,  $F_l = 0.03$ ,  $F_h = 0.4$ . For CUReT and KTH-TIPS datasets, we randomly selected  $N$  sample images from each class to construct the training sample set, and the remaining sample images were collected to construct the testing sample set in each round. This experiment was repeated 1000 times independently, and the average precision was calculated as the final recognition accuracy.

### 4.3 Experimental Results

**Optimal feature fusion weight.** In our method, the weighted average method is used to realize the global and local feature fusion, where the optimal value of weight parameter  $w$  can be set by experiment. Considering the KTH-TIPS dataset contains a large variety of scale, pose and illumination conditions, we conducted the experiments on KTH-TIPS dataset to determine the optimal value of weight  $w$ . The recognition accuracies with different values of  $w$  are listed in Table 1.

From Table 1, we could observe that: (1) when  $w = 0.7$ , the proposed method obtained the highest recognition accuracy of 99.62%, which showed the optimal fusion result of global and local features; (2) when  $w = 0$ , only the local binary pattern

**Table 1** Recognition accuracies (%) with different values of weight  $w$

Weight $w$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Recognition accuracy	98.93	99.06	99.15	99.28	99.36	99.45	99.52	99.62	99.56	99.49	98.54

**Table 2** Recognition accuracy (%) comparison

Recognition method		KTH-TIPS	CUReT
Gabor filtering based methods	Gabor filtering method [14]	91.48	94.89
	LGWBP [16]	96.87	97.66
	GLGF [17]	99.36	99.60
Texton learning based methods	VZ-Joint [10]	95.46	97.71
	VZ-MR8 [9]	93.50	97.31
	TEISF_f [11]	98.90	99.54
LBP based methods	LBP <sup>riu2</sup> [5]	92.44	95.84
	CLBP [6]	97.19	97.39
	CLBC [7]	97.03	97.16
	SSLBP [8]	99.39	99.51
	DRLBP [18]	96.78	96.41
	The proposed method	99.62	99.85

was used for texture image recognition, and the recognition accuracy decreased to 98.93%; (2) when  $w = 1$ , only the global Gabor feature was used for texture image recognition, and the recognition accuracy dropped to 98.54%. Therefore, we set  $w = 0.7$  to obtain the best fusion result of global and local texture features.

**Comparison with other methods.** In this section, we introduced more state-of-the-art methods for comparison to demonstrate the recognition performance of the proposed method, and these compared methods included methods based on Gabor filtering, texton learning and LBP. For fair comparison, we set the number of training samples per class  $N = 46$  for the CUReT dataset and  $N = 40$  for the KTH-TIPS dataset. The recognition accuracies of these methods are listed in Table 2. For other methods, we cropped the results from their original and related papers.

From Table 2, we found that: (1) the proposed method provided the best recognition accuracies of 99.85% on the CUReT dataset and 99.62% on the KTH-TIPS dataset, which outperformed all the other methods. (2) The recent LGWBP method was also a combination of LBP and Gabor filtering, which was similar to our method, but the recognition accuracy of LGWBP is only 97.66% on the CUReT dataset and 96.87% on the KTH-TIPS dataset, which is much lower than that of our method. In a word, the proposed method obtained higher recognition accuracy than other compared methods, which also demonstrated that the proposed method had superior performance in texture image recognition.

**Robustness to the number of training samples.** The robustness to the number of training samples is very important because it is difficult to acquire enough training samples in many real-world scenarios. To evaluate the robustness of the proposed method, we respectively set the number of training samples  $N = 40, 30, 20, 10$ . We

**Table 3** Recognition accuracies (%) with different  $N$ 

$N$	40	30	20	10
Gabor filtering method [14]	91.48	88.56	83.45	73.03
CLBP [6]	97.19	95.80	92.81	85.78
LGWBP [16]	96.87	95.26	91.90	83.18
CLBC [7]	97.03	95.16	92.37	85.61
GLGF [17]	99.36	98.73	97.11	91.86
The proposed method	99.62	99.26	98.50	94.82

compared the results of the proposed method with those of some other methods on KTH-TIPS dataset, and Table 3 showed the recognition accuracies.

As can be seen from Table 3, the proposed method had the smallest drop of recognition accuracy when the parameter  $N$  decreased, which showed the advantage of the proposed method in terms of robustness. The main reasons are as follows: (1) the extracted global Gabor feature and local binary pattern were discriminative, and the feature fusion of global Gabor feature and local binary pattern further improved the discriminative ability of the proposed method; (2) the image pyramid space increased the number of samples, which overcome the disadvantage of lacking enough training samples; (3) we selected three radii for local binary pattern extraction, which could capture more information with different scales.

**Efficiency evaluation.** High efficiency is very important in particular when applying the method to image sequences or large image datasets. To evaluate the efficiency of the proposed method, we compared the average running time for one texture sample among different methods, which were shown in Table 4.

As can be seen from Table 4, the proposed method had good computational efficiency. Compared with the traditional Gabor filtering method and LBP based methods, which are all efficient methods, the proposed method had comparable efficiency. Meanwhile, the time cost of the proposed method was much lower than that of texon learning based methods. Therefore, the proposed method is efficient enough in many practical applications.

**Table 4** Average running time (seconds) comparison

Recognition method	KTH-TIPS	CUReT
Gabor filtering method [14]	0.07	0.16
CLBP [6]	0.08	0.50
CLBC [7]	0.07	0.33
LGWBP [16]	0.12	0.37
GLGF [17]	0.38	0.43
VZ-MR8 [9]	4.2	7.3
TEISF_f [11]	8.0	13.6
The proposed method	0.10	0.39

## 5 Conclusion

We proposed an effective texture recognition method based on the fusion of global Gabor feature and local binary pattern. To extract discriminative and robust texture feature, an image pyramid space is constructed for global Gabor feature extraction, and multiple radii are selected for local binary pattern extraction. Finally, the global Gabor feature and local binary pattern are fused for the texture image recognition by the NSC classifier. Experimental results showed that the proposed method could achieve state-of-the-art recognition accuracy with high efficiency. Meanwhile, the proposed method is robust to scale variation and the number of training samples.

**Acknowledgements** This work was supported in part by the Scientific and Technological Project of Henan Province under Grant 202102210331, in part by the National Natural Science Foundation of China under Grant 61702462, in part by the Scientific and Technological Project of Henan Province under Grant 182102210607 and Grant 192102210108, in part by the Young and middle-aged backbone teacher program of Pingdingshan University, in part by the Ph.D. research start-up fund of Pingdingshan University under Grant PXY-BSQD-202004.

## References

- Chen, J., Chen, Z., Chi, Z., et al.: Facial expression recognition in video with multiple feature fusion. *IEEE Trans. Affect. Comput.* **9**(1), 38–50 (2018)
- Minh-Tan, P., Sébastien, L., Erchan, A.: Local feature-based attribute profiles for optical remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **56**(2), 1199–1212 (2018)
- Xie, Y., Zhang, J., Xia, Y., et al.: Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Inf. Fusion* **42**, 102–110 (2018)
- Faten, S., Ali, D.: Robust color texture descriptor for material recognition. *Pattern Recogn. Lett.* **80**, 15–23 (2016)
- Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
- Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.* **19**(6), 1657–1663 (2010)
- Zhao, Y., Huang, D.S., Jia, W.: Completed Local binary count for rotation invariant texture classification. *IEEE Trans. Image Process.* **21**(10), 4492–4497 (2012)
- Guo, Z., Wang, X., Zhou, J., et al.: Robust texture image representation by scale selective local binary patterns. *IEEE Trans. Image Process.* **25**(2), 687–699 (2016)
- Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. *Int. J. Comput. Vision* **62**(1/2), 61–81 (2005)
- Varma, M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(11), 2032–2047 (2009)
- Xie, J., Zhang, L., You, J., et al.: Effective texture classification by texton encoding induced statistical features. *Pattern Recogn.* **48**, 447–457 (2015)
- Jones, J., Palmer, L.: An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**(6), 1233–1258 (1987)
- Marcelja, S.: Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.* **70**(11), 1297–1300 (1980)

14. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**(8), 837–842 (1996)
15. Arivazhagan, S., Ganesan, L., Padam, P.S.: Texture classification using Gabor wavelets based rotation invariant features. *Pattern Recogn. Lett.* **27**, 1976–1982 (2006)
16. Hadi, H.: Multi-resolution local Gabor wavelets binary patterns for gray-scale texture description. *Pattern Recogn. Lett.* **65**, 163–169 (2015)
17. Wang, J., Fan, Y., Li, Z., et al.: Texture classification using multi-resolution global and local Gabor features in pyramid space. *SIViP* **13**(1), 163–170 (2019)
18. Rakesh, M., Karen, E.: Dominant rotated local binary patterns (DRLBP) for texture classification. *Pattern Recogn. Lett.* **71**, 16–22 (2016)

# Multi Association Semantics-Based User Matching Algorithm Without Prior Knowledge



Qiuyan Jiang and Daofu Gong

**Abstract** Cross-network user matching is one of the basic issues for realizing social network data integration. Existing research based on structure features provides a good matching method for nodes with high-degree, but ignores the matching of nodes with low values. This paper proposes an unsupervised cross-network user matching algorithm based on association semantics to solve the problem of cross-network user matching. It can effectively solve the low-degree user matching problem when the quality of the text type attributes of network users cannot be guaranteed. First, build a social network graph based on multiple types of user behaviors, and define different association semantics according to different behaviors; then, walk based on the association semantics to obtain user node sequences with association semantics, and then combine the network embedding model to learn the user feature vector representation is obtained; finally, the user similarity is measured based on the user vector similarity, and then cross-network user matching is realized. The highlights of this paper are two folds. (1) We treat the association as nodes. (2) Social network user graphs with association semantics can enrich structure feature; walks based on association semantics can obtain walks that are closer to natural language short sentences. The experimental results show that the proposed method can effectively match user with similar behavior characteristics without any prior knowledge, and can effectively achieve the matching of low-degree nodes in heterogeneous networks.

**Keywords** Cross-network user matching · Association semantics · Representation learning · User behavior analysis

---

Q. Jiang · D. Gong ()

State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou

450001, China

e-mail: [dfgong@aliyun.com](mailto:dfgong@aliyun.com)

Q. Jiang

College of Science and Information, Qingdao Agricultural University, Qingdao 266001, China

## 1 Introduction

Social networking has become an indispensable part of people's lives due to the development and popularization of network technology and network services. Research on user profile problems on social networks has also aroused widespread concern among researchers, which is to achieve the labeling of the users based on social network user's data. It is the basis of public opinion monitoring and accurate recommendation systems, etc. At present, the research on social network user profiles is mostly based on a single social network, which limits its application. In order to effectively enrich the research area and to reduce the limitations of a single social network, a better choice is to fuse the different social networks. However, the major premise of the cross-network data fusion problem is to correctly match the user accounts belonging to the same natural user across different networks, that is, the cross-network user matching. As a result, the research of cross-network user matching has gained high academic attention and significance.

The user-related data generated by network users using the network platform is the data basis for studying the problem of cross-network user matching. According to the data type, user matching methods are usually divided into three categories: matching methods based on text attributes [1–10] based on structure features [11–16] and based on comprehensive attributes. The text mainly includes screen name, gender, birthday, city and profile image and user-generated content, such as the content posted by users on the network, posting time, location and writing style, etc., while the network structure mainly refers to the association between users. The matching effect of the text-based method depends on the quality of the attribute data. The associated structure data implies the user's real dynamic behavior characteristics, and these behavior characteristics usually have a certain degree of stability and are not easy to be imitated. In addition, most of the existing algorithms solve the social network user matching problem based on the known anchor nodes. Among them, the anchor nodes of most algorithms are determined based on the matching of user text attributes. However, the quality of text attributes is indeed difficult to evaluate. Therefore, it is particularly important to find anchor nodes based on the similarity of the associated structure to achieve user matching. Therefore, the user-associated structure is a significant factor to consider. However, the existing methods are usually based on homogeneous network research, but do not fully consider the characteristics of heterogeneous networks. This is also the reason why low-degree nodes are difficult to match.

In order to solve the above problems, this paper proposes a cross-network user matching algorithm based on association semantics without Prior Knowledge. First, build a social network graph based on multiple types of user behaviors, and define different association semantics according to different types of user behaviors; then, walk based on the association semantics to obtain node sequences with association semantics, and then combine the network embedding model to learn the user's feature vector representation; finally, the user similarity is measured based on the user vector

similarity, and then cross-network user matching is realized. The proposed method is without Prior Knowledge, and the contributions are as follows:

1. It proposes to define different association semantics according to different types of behavior;
2. The improved walking model is a walking model based on association semantics, and the walking sequence has semantic walking sequence;
3. Effectively realize the representation and matching of low-level nodes. The results show that the effect in the user matching task is good.

This article proceeds as follows. Section 2 reviews the related works on cross-network user matching. Section 3 analyzed the algorithm theoretically and described the proposed MASUM-P algorithm in detail. Section 4 covers the experimental results. And Sect. 5 concludes the overall paper.

## 2 Related Works

The user matching algorithm based on structure feature is developed based on the idea that user has similar structures in different social networks. Such algorithms are often based on seed nodes, and then use structure features to achieve user matching [11–13, 17, 18]. The quality of the seed nodes directly affects the effect of the algorithm. In recent years, network representation learning algorithms have attracted people's attention, and they have been combined with methods such as co-training and iterative processes to address user matching problems [16, 19, 20]. [14, 21] develop a framework named IONE to model the follower-ship and followee-ship as input and output context vectors. [22] propose an attention-based network embedding model. [23, 24] are both multi-view representation learning combined with text attributes. [16] propose Friend-based User Identification without Prior Knowledge (FRUI-P). This method uses a random walk method to generate user node sequences. Based on the representation learning method, the user node is represented as a low-dimensional vector, and the Euclidean distance of the vector is used to measure the similarity of the vector to achieve user matching.

However, existing researches on homogeneous networks treat the associations between users equally, ignoring the differences in association semantics. Therefore, this paper defines association semantics and proposes a cross-network user matching algorithm based on association semantics to effectively improve the matching effect of heterogeneous network users.

### 3 Multi Association Semantics-Based User Matching Algorithm Without Prior Knowledge(MASUM-P)

The proposed cross-network user matching algorithm based on user association semantics will be described in this section.

#### 3.1 Motivation

Existing research basically builds a user relationship graph based on the “follow” or “friend” relationship between users. However, the “follow” or “friend” relationship may only indicate a state between users at a certain moment. For example, it is possible that user  $a$  and user  $b$  have established a “follow” or “friend” relationship and then cancel each other’s “follow” or “friend”. Therefore, this relationship is inherently unstable. In addition, some users may not have a “follow” or “friend” relationship at all due to privacy protection and other factors. Such users will not be included in the user relationship graph, and it is impossible to achieve user matching. Therefore, the social network graph based on the “follow” or “friend” relationship is actually just a subgraph of the social network that has lost other types of user relationships. In fact, users in social networks have more complex and diverse behaviors. Different user behaviors correspond to user relationships with different semantics, and different relationships collectively reflect the state of user relationships in a period of time.

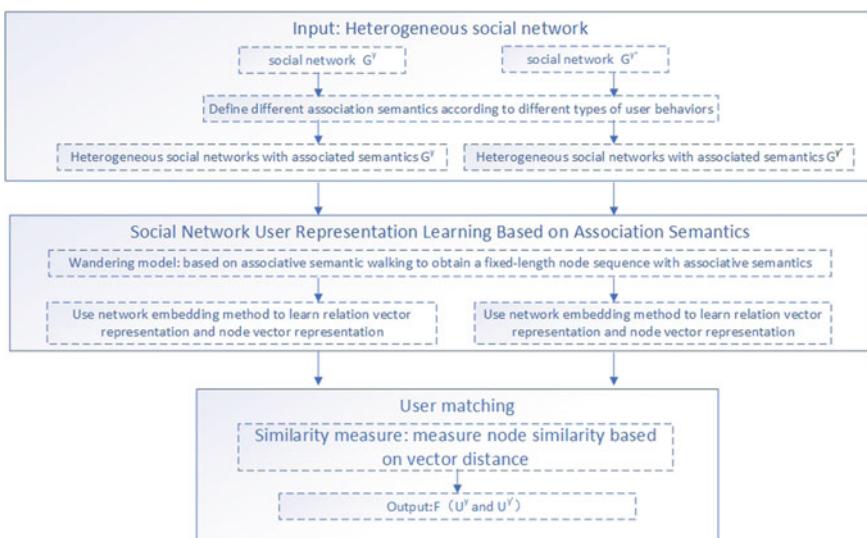
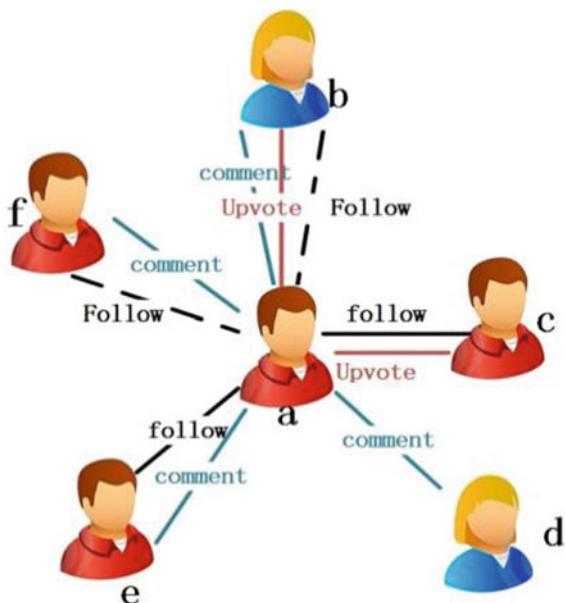
Therefore, this paper will establish heterogeneous social networks based on multiple types of user relationships. Common types of relationships between users in social networks include “follow”, “comment”, “like”, “mention”, “repost”, etc.... No matter which one, different types represent different association semantics. The association semantics are just like the “verbs” in natural language, which meaningfully link user nodes.

As shown in Fig. 1, assigning different association semantics based on different types of user behaviors effectively enriches the characteristics of users.

#### 3.2 Model Framework

Figure 2 shows the overview of the proposed algorithm. The algorithm consists of three modules: the social network representation module with association semantics, the social network user representation learning module based on association semantics, and the user matching module. The social network representation module will define association semantics based on different types of behaviors of users, and construct a social network graph with semantic. The social network user representation learning module based on association semantics will walk based on the associated semantics, obtain the node sequence with the associated semantics and

**Fig. 1** A multi-type and multi-relational social network



**Fig. 2** Overview of MASUM-P

use it as the input of the network embedded learning model, and then obtain the feature vector representation of the user and the association.

### 3.3 Graph Representation for Social Network

Here is an observation in the social networks that there exist different types of associations among users. Due to the importance of user behavior characteristics, a graph for social networks could be built considering all behaviors among users. Different types of behaviors represent different associated semantics. Usually, social networks are represented as a graph, defined as follows:

Definition 1: social networks graph.

The social network graph is represented as  $G = \{U, E, X\}$ , where  $U = \{u_m | m = 1, \dots, |U|\}$  is the user set of the network  $G$ .  $X$  is the set of association types,  $X = \{x | x = 1, \dots, |X|\}$ . That is, there are  $|X|$  types of association in the social network  $G$ .  $E$  is the association set between users.  $E = \{e_{u_m u_n}^x | x \in X, u_m, u_n \in U\}. e_{u_m u_n}^x = 1$ , it indicates that two users are associated with  $x$ -type.

### 3.4 Network User Representation Based on User Association Semantics

The representation learning of network users is the key to solve the user matching problem, whose purpose is to represent each social network user  $u_m$  as a real feature vector  $v_m$ ,  $v_m \in V$  with dimension  $d$  ( $d \ll |U|$ ).

It is the key to solve the network user representation to construct a user node sequence based on association semantics. This paper adopts a basic random walk model: assuming that the current node is *node* and the neighbor node is represented as *Neighbor(node)*, then the sampling probability of the target node *Target\_node* can be represented as:

$$P(\text{Target\_node} | \text{node}) = \begin{cases} \frac{1}{|\text{Neighbor}(\text{node})|}, & \text{Target}_\text{node} \in \text{Neighbor}(\text{node}) \\ 0, & \text{else} \end{cases} \quad (1)$$

Each user node in the social network should be set as the current node and repeat the above process to get a walk sequence according to sampling probability. After generating all nodes sequence, node feature vectors are learned by word2vec [25] model which takes node sequences as input.

Algorithm 1 shows the network user representation learning algorithm based on user association semantics.

**Algorithm:** User representation learning based on user association semantics

Input	$G = \{U, E, X\}$
	Parameters for walk ( $l, t$ ),
	Parameters for representation learning ( $d$ , window size, $sg$ , $hs$ )
Output	Network representation matrix $V_U = \{v_{u_m}, m = 1, \dots,  U \}$ $V_X = \{v_{x_i}, i = 1, \dots,  X \}$
Step1	Determine the different types of relationships that the current user exists according to the given original network data, and determine the association semantics according to different types of behaviors
Step2	Each user node serves as the starting node and the current node, and the transfer target is determined according to the sampling probability. After $(l - 1)/2$ walks, a node sequence of length $l$ ( $l$ odd number) is generated
Step3	Repeat step 2, and each user node is used as the starting node to walk $t$ times to obtain all the walking sequences, which are used as the input of the network embedding model
Step4	Use the network embedding model to learn to obtain all the node representation vectors
Step5	Output $V_U, V_X$

### 3.5 Cross-Network User Matching

This section introduces user matching criteria based on the similarity of user feature vectors. The similarity between the user feature vectors can be measured according to the vector distance. It is generally believed that the closer the distance, the greater the similarity. The user node pair with the highest score of similarity is regarded as the user matching result.

Vector distance can be measured in different ways, such as Euclidean Distance, Chebyshev distance, etc. This paper uses Euclidean Distance to measure the vector distance. Given the social networks  $G^y$  and  $G^{y'}$  to be matched, the feature vectors of the user  $u_m^y \in U^y$  and the user  $u_n^{y'} \in U^{y'}$  are  $v_m^y$  and  $v_n^{y'}$ . Then, Euclidean Distance of  $v_m^y$  and  $v_n^{y'}$  is defined as:

$$D(v_m^y, v_n^{y'}) = \sqrt{(v_m^y - v_n^{y'})(v_m^y - v_n^{y'})^T} \quad (2)$$

The similarity  $s(u_m^y, u_n^{y'})$  between  $u_m^y$  and  $u_n^{y'}$  is standardized as:

$$s(u_m^y, u_n^{y'}) = \frac{1}{1 + \log[D(v_m^y, v_n^{y'}) + 1]} \quad (3)$$

The similarity between the user  $u_m^y$  in the social network  $G^y$  and each user in the social network  $G^{y'}$  is expressed as  $S_m = \{s(u_m^y, u_n^{y'}) \mid u_n^{y'} \in U^{y'}, n = 1 \dots |U^{y'}|\}$ . The user pair with the largest similarity value  $\arg \max(S_m)$  is regarded as the matching result.

## 4 Experiments

This section introduces the experimental datasets, evaluation methods and experimental results.

### 4.1 Experimental Dataset and Settings

As far as we know, there is no consistent benchmark dataset for user matching task across social networks. We crawls some user data from the “ZhiHu” network, the number of users is 95,716. The user characteristic data is extracted from the crawled data, which mainly includes user nodes and behavioral associations between users, and the degree distribution of the graph obeys a power law distributed. This paper constructs the synthetic datasets [12]. These two parts of data are  $G^y$  and  $G^{y'}$ . The numbers of nodes are 6883 and 6920 respectively; the node overlap rate is 68.6%. Each node contains 1–4 different types of associations.

The length of the walk is 15, 31 and 51, the number of walks of each node is 100, 300 and 500, the sliding window is 10, and the dimension of the feature vector is 300 respectively. Use random walk.

### 4.2 Evaluation Metrics

With regard to the user matching problem, the precision represents how many of the samples predicted to be matched users are actually matched users. Therefore, to evaluate the matching result we put more emphasis on matching precision. The larger the precision value, the better the algorithm performance. Definitions [16] are shown:

$$\text{Precision} = \frac{\text{Number of correctly matched users}}{\text{Total number of matched users}} \quad (4)$$

**Table 1** The impact of associative semantics on matching results

Associative semantic types	Proportion of nodes that match correctly and contain more than two types of associations	
	$I = 15, t = 100$	$I = 15, t = 300$
2	80.1%	80.2%
3	83.3%	85.6%
4	72%	70%

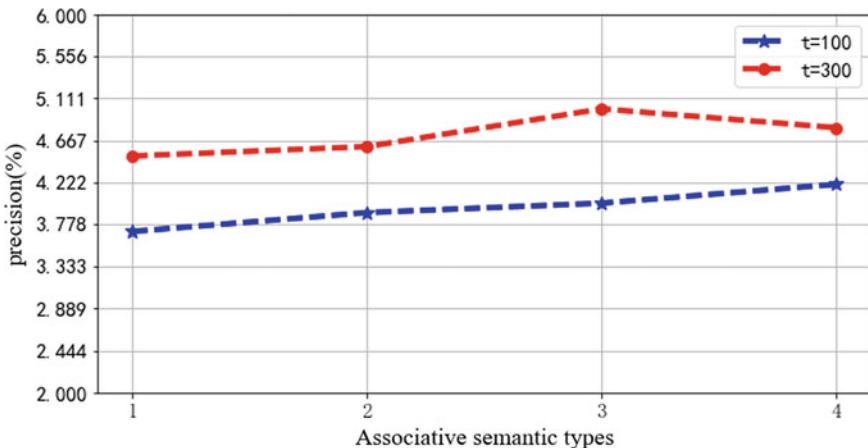
### 4.3 Experimental Results and Analysis

We studied the impact of user association semantics on matching results. As the number of association semantics included in the network to be matched increases, the network becomes more and more complex. When the types of association semantics included in the network to be matched increase, the accuracy of user matching gradually increases. Moreover, averages of 78.5% of correctly matched user nodes have more than two types of relational semantics. It shows that the association semantics defined based on user network behavior is helpful to improve the matching effect of users.

The existing user matching algorithms based on structural features are extremely unfriendly to low-degree nodes (the degree value is less than or equal to 6). Because, relative to the height-degree node, the node's degree value is low, and the data representing the characteristics is lacking, so the low-degree value node is often removed. The method can solve this problem well after the introduction of relational semantics. The results of many experiments show that there are 89.9% of correctly matched user nodes with low-degree on average. Because the introduction of association semantics increases the structural characteristics of degree nodes, the representation vector of such nodes can be better learned, so as to better realize the matching of low-degree nodes. Table 1 and Fig. 3 show the matching results.

## 5 Conclusion

The purpose of cross-network user matching is to accurately match accounts belonging to the same natural user on different network platforms, thereby realizing the integration of network data and laying a data foundation for complex network analysis problems. This paper proposes an unsupervised cross-network user matching algorithm based on association semantics based on the same user having similar behavior characteristics in different networks. The algorithm constructs a social network graph based on association semantics and constructs a sequence of nodes with semantically related “context” based on the associated semantic walk method, which lays a good data foundation for the representation learning method. This method does not require prior knowledge, so the algorithm is not affected



**Fig. 3** The impact of associative semantics on precision

by the quality of text attribute information. The algorithm makes full use of user behavior characteristics to provide a matching method for users with low-degree in the network.

## References

- Marco, B., Christian, P., Thorsten, H., Engin, K., Davide, B., Christopher, K.: Abusing social networks for automated user profiling. In: International Workshop on Recent Advances in Intrusion Detection, Berlin, Heidelberg, pp. 422–441 (2010)
- Daniele, P., Claude, C., Mohamed, A.K., Pere, M.: How unique and traceable are usernames? In: International Symposium on Privacy Enhancing Technologies Symposium, Berlin, Heidelberg, pp. 1–17 (2011)
- Elie, R., Richard, C., Albert, D.: User profile matching in social networks. In: 2010 13th International Conference on Network-Based Information Systems, Japan, pp. 297–304 (2010)
- Jing, L., Fan, Z., Xinying, S., Young-In, S., Chin-Yew, L., Hsiao-Wuen, H.: What's in a name? An unsupervised approach to link users across communities. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, New York, NY, United States, pp. 495–504 (2013)
- Kaikai, D., Ling, X., Langshui, Z., Honghai, W., Ping, X., Feifei, G.: A user identification algorithm based on user behavior analysis in social networks. IEEE Access **7**, 47114–47123 (2019)
- Yongjun, L., Zhen, Z., You, P., Hongzhi, Y., Quanqing, X.: Matching user accounts based on user generated content across social networks. Futur. Gener. Comput. Syst. **83**, 104–115 (2018)
- Yongjun, L., You, P., Wenli, J., Zhen, Z., Quanqing, X.: User identification based on display names across online social networks. IEEE Access 1319–1326 (2017)
- Rong, Z., Jiexun, L., Hsinchun, C., Zan, H.: A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inform. Sci. Technol. **57**, 378–393 (2006)
- Mishari, A., Gene, T.: Exploring linkability of user reviews. In: European Symposium on Research in Computer Security, Berlin, Heidelberg, pp. 307–324 (2012)

10. Chris, R., Yunsung, K., Augustin, C., Nitish, K., Silvio, L.: Linking users across domains with location data: theory and validation. In: Proceedings of the 25th International Conference on World Wide Web, Canada, pp. 707–719 (2016)
11. Shulong, T., Ziyu, G., Deng, C., Xuzhen, Q., Jiajun, B., Chun, C.: Mapping users across networks by manifold alignment on hypergraph. In: Twenty-Eighth AAAI Conference on Artificial Intelligence, Canada, pp. 159–165 (2014)
12. Zhongbao, Z., Qihang, G., Tong, Y., Sen, S.: Identifying the same person across two similar social networks in a unified way: globally and locally. *Inf. Sci.* **394**, 53–67 (2017)
13. Nitish, K., Silvio, L.: An efficient reconciliation algorithm for social networks. In: Proceedings of the VLDB Endowment, vol. 7, pp. 377–388 (2014)
14. Li, L., William, K.C., Xin, L., Lejian, L.: Aligning users across social networks using network embedding. In: IJCAI2016, New York, USA, pp. 1774–1780 (2016)
15. Wen, Z., Kai, S., Huang, L., Yalin, W.: Graph neural networks for user identity linkage. arXiv preprint [arXiv:1903.02174](https://arxiv.org/abs/1903.02174) (2019)
16. Xiaoping, Z., Xun, L., Xiaoyong, D., Jichao, Z.: Structure based user identification across social networks. *IEEE Trans. Knowl. Data Eng.* **30**(6), 1178–1191 (2018)
17. Xiaoping, Z., Xun, L., Haiyan, Z., Yuefeng, M.: Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Trans. Knowl. Data Eng.* **28**(2), 411–424 (2016)
18. Wei, Z., Shulong, T., Ziyu, G., et al.: Learning to map social network users by unified manifold alignment on hypergraph. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(12), 5834–5846 (2018)
19. Mark, H., Haoming, S., Tara, S., Danai, K.: REGAL: representation learning-based graph alignment. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, New York, NY, USA, pp. 117–126 (2018)
20. Zexuan, Z., Yong, C., Mu, G., Zaiqing, N.: Colink: an unsupervised framework for user identity linkage. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
21. Li, L., Xin, L., William, K.C., Lejian, L.: Structural representation learning for user alignment across social networks. *IEEE Trans. Knowl. Data Eng.* **32**(9), 1824–1837 (2020)
22. Li, L., Youmin, Z., Shun, F., Fujin, Z., Jun, H., Pu, Z.: ABNE: an attention-based network embedding for user alignment across social networks. *IEEE Access* **7**, 23595–23605 (2019)
23. Yi-Yu, L., Jennifer, N.: MERL: multi-view edge representation learning in social networks. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 675–684. Association for Computing Machinery, New York, NY, USA (2020)
24. Weiqing, W., Hongzhi, Y., Xingzhong, D., Wen, H., Yongjun, L., Quoc, V.H.N.: Online User representation learning across heterogeneous social networks. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’19), pp. 545–554. Association for Computing Machinery, New York, NY, USA (2019)
25. Tomas, M., Kai, C., Greg, C., Jeffrey, D.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR (2013)

# Robust Template Matching via Hierarchical Convolutional Features from a Shape Biased CNN



Bo Gao and Michael W. Spratling

**Abstract** Finding a template in a search image is an important task underlying many computer vision applications. Recent approaches perform template matching in a deep feature-space, produced by a convolutional neural network (CNN), which is found to provide more tolerance to changes in appearance. In this article, we investigate whether enhancing the CNN’s encoding of shape information can produce more distinguishable features, so as to improve the performance of template matching. This investigation results in a new template matching method that produces state-of-the-art results in a standard benchmark. To confirm these results, we also create a new benchmark and show that the proposed method also outperforms existing techniques on this new dataset. Our code and dataset is available at: <https://github.com/iminfine/Deep-DIM>.

**Keywords** Template match · Convolutional neural networks · VGG19

## 1 Introduction

Template matching is a technique to find a rectangular region of an image that contains a certain object or image feature. It is widely used in many computer vision applications such as object tracking [1, 2], object detection [3, 4] and 3D reconstruction [5, 6]. A similarity map is typically used to quantify how well a template matches each location in an image. Traditional template matching methods calculate the similarity using a range of metrics such as the normalised cross-correlation (NCC), the sum of squared differences (SSD) or the zero-mean normalised cross correlation (ZNCC) applied to pixel intensity or color values. However, because these methods rely on comparing the values in the template with those at corresponding locations in the image patch they are sensitive to changes in lighting conditions, non-rigid deformations of the target object, or partial occlusions, which results in a low similarity score when one or multiple of these situations occur.

---

B. Gao · M. W. Spratling

Department of Informatics, King’s College London, London, UK

e-mail: [bo.gao@kcl.ac.uk](mailto:bo.gao@kcl.ac.uk)

With the help of deep features learned from convolutional neural networks (CNNs), vision tasks such as image classification [7, 8], object recognition [9, 10], and object tracking [1, 2] have recently achieved great success. In order to succeed in such tasks, CNNs need to build internal representations that are less affected by changes in the appearance of objects in different images. To improve the tolerance of template matching methods to changes in appearance recent methods have been successfully applied to a feature-space produced by the convolutional layers of a CNN [11–15].

The higher layers of CNNs are believed to learn representations of shapes from low-level features [16]. However, a recent study [17] demonstrated that ImageNet-trained CNNs are biased toward making categorisation decisions based on texture rather than shape. This work also showed that CNNs could be trained to increase sensitivity to shape and that this would improve accuracy and robustness both of object classification and detection. Assuming that shape information is also useful for template matching, these results suggest that the performance of template matching methods applied to CNN generated feature-spaces could potentially also be improved by training the CNN to be more sensitive to shape.

In this article we verified the assumption by comparing the features from four CNN models with the same network structure but differing in shape sensitivity. Our results show that training a CNN to learn about texture while biasing it to be more sensitive to shape information, can improve template matching performance. Furthermore, by comparing template matching performance when using feature-spaces created from all possible combinations of one, two and three convolutional layers of the CNN it was found that the best results were produced by combining features from both early and late layers. Early layers of a CNN encode lower-level information such as texture, while later layers encode more abstract information such as object identity. Hence, both sets of results suggest that a combination of texture and shape information is beneficial for template matching.

Our main contributions are summarized as follows: (1) We created a new benchmark which, compared to the existing standard benchmark, is more challenging, provides a far larger number of images pairs, and is better able to discriminate the performance of different template matching methods. (2) By training a CNN to be more sensitive to shape information and combining features from both early and late layers, we created a feature-space in which the performance of most template matching algorithms is improved. (3) Using this feature-space together with an existing template matching method, DIM [18], we obtained state-of-art results on both the standard and new datasets.

## 2 Related Work

To overcome the limitations of classic template matching methods, many approaches [11, 12, 14, 15, 18] have been developed. These methods can be classified into two main categories.

## 2.1 *Matching*

One category changes the computation that is performed to compare the template to the image to increase tolerance to changes in appearance. For example, Best-Buddies Similarity (BBS) counts the proportion of sub-regions in the template and the image patch that are Nearest-Neighbor (NN) matches [14]. Similarly, Deformable Diversity Similarity (DDIS) explicitly considers possible template deformation and uses the diversity of NN feature matches between a template and a potential matching region in the search image [15]. The Divisive Input Modulation (DIM) algorithm [18] extracts additional templates from the background and lets the templates compete with each other to match the image. Specifically, this competition is implemented as a form of probabilistic inference known as explaining away [19, 20] which causes each image element to only provide support for the template that is the most likely match. Previous work has demonstrated that DIM, when applied to color feature-space, is more accurate at identifying features in an image compared to both traditional and recent state-of-the-art matching methods [18].

## 2.2 *Features*

The second category of approaches changes the feature-space in which the comparison between the template and the image is performed. The aim is that this new feature-space allows template matching to be more discriminative while also increasing tolerance to appearance changes. Co-occurrence based template matching (CoTM) transforms the points in the image and template to a new feature-spaced defined by the co-occurrence statistics to quantify the dissimilarity between a template to an image [12]. Quality-aware template matching (QATM) is a method that uses a pretrained CNN model as a feature extractor. It learns a similarity score that reflects the (soft-) repetitiveness of a pattern using an algorithmic CNN layer [11].

## 2.3 *Deep Features*

Many template matching algorithms from the first category above, can be applied to deep features as well as directly to color images. The deep features used by BBS, CoTM and QATM are extracted from two specific layers of a pre-trained VGG19 CNN [21], conv1-2 and conv3-4. Following the suggestion in [2] for object tracking, DDIS also takes features from a deeper layer: fusing features from layers conv1-2, conv3-4 and conv4-4. [13] proposed a scale-adaptive strategy to select a particular individual layer of a VGG19 to use as the feature-space according to the size of template. In each case using deep features was found to significantly improve template matching performance compared to using color features.

A recent study showed that ImageNet-trained CNNs are strongly biased towards recognising textures rather than shapes [17]. This study also demonstrated that the same standard architecture (ResNet-50 [22]) that learns a texture-based representation on ImageNet is able to learn a shape-based representation when trained on ‘Stylized-ImageNet’: a version of ImageNet that replaces the texture in the original image with the style of a randomly selected painting through AdaIN style transfer [23]. This new shape-sensitive model was found to be more accurate and robust in both object classification and detection tasks. Inspired by the findings, in this paper we investigate if enhancing the shape sensitivity of a CNN can produce more distinguishable features that improve the performance of template matching.

### 3 Methods

Previous work on template matching in deep feature-space (see Sect. 2) has employed a VGG19 CNN. To enable a fair comparison with those previous results, we also used the VGG19 architecture. However, we used four VGG19 models that differed in the way they were trained to encode different degrees of shape selectivity by the same approach used by Geirhos et al. [17] (as summarised in Table 1). Model-A was trained using the standard ImageNet dataset [21] (we used the pretrained VGG19 model from the PyTorch torchvision library) which has the least shape bias. Model-B was trained on the Stylized-ImageNet dataset and thus has the most shape bias. Model-C was trained on a dataset containing the images from both ImageNet and Stylized-ImageNet. Model-D was initialised with the weights of Model-C and then fine-tuning on ImageNet for 60 epochs using a learning rate of 0.001 multiplied by 0.1 after 30 epochs. Therefore Model-C and Model-D have intermediate levels of shape bias, with model-D being less selective to shape than Model-C. The learning rate was 0.01 multiplied by 0.1 after every 30 epochs for Model-B and after every 15 epochs for Model-C. Number of epochs was 90 for Model-B and 45 for Model-C (as the dataset used to train Model-C was twice as large as that used to train Model-B the number of weight updates was the same for both models). The other training hyperparameters used for each model were: batch size 256, momentum 0.9, and weight decay 1e-4. The optimizer was SGD.

In color feature-space the DIM algorithm was previously found to produce the best performance (see Sect. 2). We therefore decided to use this algorithm to determine the best CNN feature-space to use for template matching. The DIM algorithm was

**Table 1** Four different VGG19 CNN models used in this paper. IN and SIN are the abbreviations of ImageNet and stylized-ImageNet respectively

Name	Training	Fine-tuning	Rank of shape sensitivity
Model-A	IN	–	4
Model-B	SIN	–	1
Model-C	IN + SIN	–	2
Model-D	IN + SIN	IN	3

applied to deep features in exactly the same way that it was previously applied to color images [18], except: (1) five (rather than four) additional templates were used; and, (2) the positive and rectified negative values produced by a layer of the CNN were directly separated into two parts and used as separate channels for the input to the DIM algorithm (in contrast, previously each channel of a color image was processed using a difference of Gaussians filter and the positive and rectified negative values produced were used as separate channels for the input to the DIM algorithm).

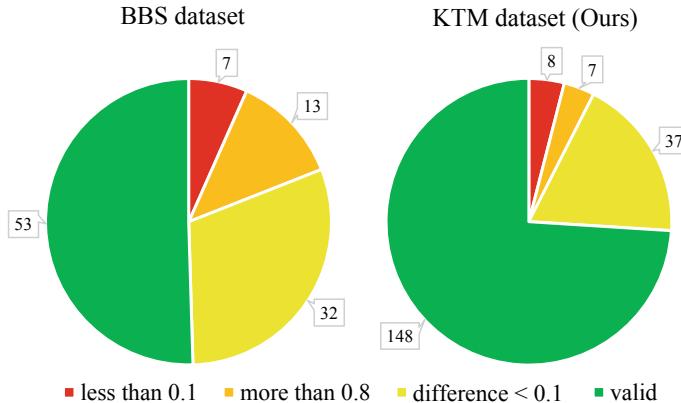
## 4 Results

### 4.1 Dataset Preparation

The BBS dataset [14] has been extensively used for the quantitative evaluation of template matching algorithms [11–15]. This dataset contains 105 template-image pairs which are sampled from 35 videos (3 pairs per video) from a tracking dataset [24]. Each template-image pair are taken from frames of the video that are 20 frames apart. To evaluate the performance of a template matching algorithm the intersection-over-union (IoU) is calculated between the predicted bounding box and the ground truth box for the second image in the pair. The overall accuracy is then determined by calculating the area under the curve (AUC) of a success curve produced by varying the threshold of IoU that counts as success.

Although the BBS data is widely used, it is not particularly good at discriminating the performance of different template matching methods. To illustrate this issue we applied one baseline method (ZNCC) and three state-of-art methods (BBS, DDIS and DIM) to the BBS dataset in color space. The results show that there are 52 template-image pairs where all methods generate very similar results: these can be sub-divided into 7 template-image pairs for which all methods fail to match (IoU less than 0.1 for all four methods), 13 template-image pairs for which all methods succeed (IoU greater than 0.8 for all four methods), and 32 template-image pairs for which all methods produce similar, intermediate, IoU values within 0.1 of each other. This means that only 53 template-image pairs in the BBS dataset help to discriminate the performance of these four template matching methods. These results are summarised in Fig. 1.

We, therefore, created a new dataset, the King’s Template Matching (KTM) dataset, following a similar procedure to that used to generate the BBS dataset. The new dataset contains 200 template-image pairs sampled from 40 new videos (5 pairs per video) selected from a different tracking dataset [25]. In contrast to the BBS dataset, the template and the image were chosen manually to avoid pairs that contain significant occlusions and non-rigid deformations of the target that no method is likely to match successfully, and the image pairs were separated by 30 (rather than 20) frames to reduce the number of pairs for which matching would be easy for all methods. These changes make the new data more challenging and provide a far



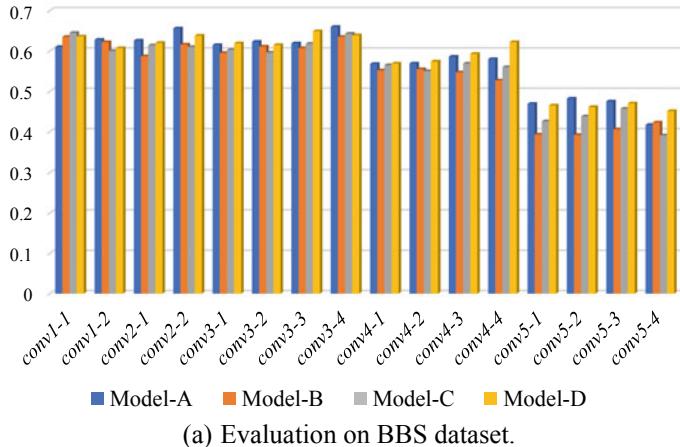
**Fig. 1** Discriminative ability of two datasets evaluated by comparing the IoU scores produced by ZNCC, BBS, DDIS and DIM

larger number of images pairs that can discriminate the performance of different methods, as shown in Fig. 1. Both the new dataset and the BBS dataset were used in the following experiments.

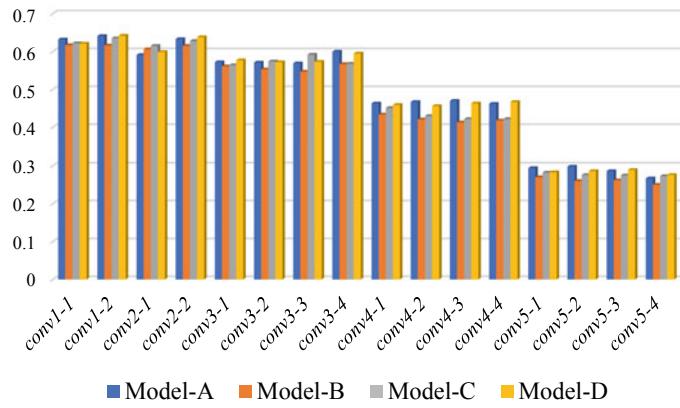
#### 4.2 *Template Matching Using Features from Individual Convolutional Layers*

To reveal how the shape bias affects template matching, we calculate AUC using DIM with features from every single convolutional layer of the four models. As the features from the later convolutional layers are down-sampled using maxpooling, by a factor of  $\frac{1}{2}$ ,  $\frac{1}{4}$ ,  $\frac{1}{8}$  and  $\frac{1}{16}$  compared to the original image, the bounding box of the template is also multiplied by the same scaling factor and the resulting similarity map is resized back to the original image size to make the prediction. The AUC scores across the BBS and KTM datasets are summarised in Fig. 2.

For all four models there is a tendency for the AUC to be higher when template matching is performed using lower layers of the CNN compared to later layers. This suggests that template matching relies more on low-level visual attributes, such as texture, rather than higher-level ones such as shape. Among the three models trained with Stylized-ImageNet, the AUC score for most CNN layers is greater for Model-D than Model-C, and greater for Model-C than Model-B. This also suggests that template matching relies more on texture features than shape features. Comparing Model-A and Model-D, it is hard to say which one is better. However, the AUC score calculated on the BBS dataset using features from *conv4-4* of Model-D is noticeably better than that for Model-A. This suggests that increasing the shape-bias of later layers of the CNN could potentially lead to better template matching. However, these results are not reflected by the results for the KTM dataset. One possible explanation



(a) Evaluation on BBS dataset.



(b) Evaluation on KTM dataset.

**Fig. 2** The AUC scores of DIM using features from different convolutional layers of four models

is that the templates in the KTM dataset are smaller, in general, than those in the BBS dataset (if the template size is defined as the product of its width and height, then the mean template size of for the KTM datasets 1603 whereas it is 3442 for the BBS dataset). Smaller templates tend to be less discriminative. The sub-sampling that occurs in later levels of the CNN results in templates that are even smaller and less discriminative. This may account for the worse performance of the later layers of each CNN when tested using the KTM dataset rather than the BBS dataset. It is also a confounding factor in attributing the better performance of the early layers to a reliance on texture information.

### 4.3 Template Matching Using Features from Multiple Convolutional Layers

We compared Model-A and Model-D by applying the DIM template matching algorithm to features extracted from multiple convolutional layers of each CNN. To combine feature maps with different sizes bilinear interpolation was used to make them the same size. If the template was small (height times width less than 4000) the feature maps from the later layer(s) were scaled to be the same size as those in the earlier layer(s). If the template was large, the feature maps from the earlier layer(s) were reduced in size to be the same size as those in the later layer(s). To balance low and high level features, the dimension of the features maps from the latter layer(s) were reduced by PCA to the same number as those in the earlier layer.

Table 2 shows the AUC scores produced by DIM using features from two convolutional layers of Model-A and Model-D. All possible combinations of two layers were tested, and the table shows only selected results with the best performance.

**Table 2** Partial AUC scores of DIM using features from two convolutional layers of model-A (upper value in each cell) and Model-D (lower value in each cell)

(a) Evaluation on BBS dataset

Layer	Layer					
	conv3-3	conv3-4	conv4-1	conv4-2	conv4-3	conv4-4
	AUC					
conv1-1	0.710 0.707↓	0.705 0.714↑	0.713 0.704↓	0.697 <b>0.718</b> ↑	0.698 0.710↑	0.711 0.708↓
conv1-2	0.686 0.686	0.686 0.687↑	0.674 0.707↑	0.655 0.696↑	0.680 0.690↑	0.683 0.710↑
conv2-1	0.658 0.659↑	0.670 0.669↓	0.664 0.665↑	0.653 0.671↑	0.662 0.683↑	0.667 0.693↑
conv2-2	0.659 0.665↑	0.661 0.667↑	0.653 0.676↑	0.641 0.679↑	0.659 0.676↑	0.663 0.682↑

(b) Evaluation on KTM dataset

Layer	Layer					
	conv3-3	conv3-4	conv4-1	conv4-2	conv4-3	conv4-4
	AUC					
conv1-1	0.687 0.689↑	0.684 0.691↑	0.677 0.682↑	0.668 0.695↑	0.670 0.684↑	0.678 0.687↑
conv1-2	0.687 0.680↓	0.689 0.694↑	0.682 0.695↑	0.685 <b>0.697</b> ↑	0.675 0.691↑	0.682 0.690↑
conv2-1	0.634 0.642↑	0.633 0.651↑	0.647 0.665↑	0.645 0.671↑	0.655 0.668↑	0.639 0.666↑
conv2-2	0.642 0.657↑	0.651 0.664↑	0.664 0.670↑	0.661 0.673↑	0.669 0.669	0.669 0.669

**Table 3** Best 10 results when using combinations of features from three convolutional layers of Model-D.  $C12_{44}^{41}$  means fusing the features from  $conv1\text{-}2$ ,  $conv4\text{-}1$  and  $conv4\text{-}4$  for instance

(a) Evaluation on BBS dataset										
Layers	$C12_{44}^{41}$	$C11_{43}^{34}$	$C11_{44}^{42}$	$C12_{52}^{43}$	$C11_{43}^{22}$	$C11_{44}^{41}$	$C11_{43}^{41}$	$C11_{42}^{34}$	$C12_{44}^{43}$	$C11_{44}^{34}$
AUC	0.728	0.727	0.724	0.724	0.723	0.722	0.720	0.720	0.720	0.720

(b) Evaluation on KTM dataset										
Layers	$C11_{42}^{34}$	$C12_{43}^{32}$	$C11_{43}^{34}$	$C12_{42}^{22}$	$C12_{42}^{31}$	$C12_{42}^{34}$	$C12_{43}^{34}$	$C12_{43}^{31}$	$C12_{43}^{33}$	$C11_{42}^{33}$
AUC	0.711	0.709	0.708	0.706	0.706	0.705	0.705	0.705	0.705	0.704

Each cell of the table contains two AUC scores, the upper one is produced using Model-A and the bottom is produced by Model-D. The up and down-arrows indicate whether the AUC score of Model-D is better or worse than that of Model-A.

It can be seen from Table 2 that for the 24 layer combinations for which results are shown, 21 results for both BBS and KTM dataset are better for Model-D than for Model-A. Furthermore, the best result for each dataset (indicated in bold) is generated using the features from Model-D. These results thus support the conclusion that more discriminative features can be obtained by slightly increasing the shape bias of the VGG19 model which increases the performance of template matching.

To determine if fusing features from more layers would further improve template matching performance, DIM was applied to all combinations of three layers from Model-D. There are a total of 560 different combinations of three layers. It is impossible to show all these results in this paper, therefore the highest 10 AUC scores are shown in Table 3. For both datasets, using three layers produced an improvement in the best AUC score (around 0.01) compared to when using two layers.

#### 4.4 Comparison with Other Methods

This section compares our results with those produced by other template matching methods in both color and deep feature-space. When evaluated on the BBS dataset, the deep features used by each template matching algorithm were the features from layers  $conv1\text{-}2$ ,  $conv4\text{-}1$  and  $conv4\text{-}4$  of Model-D. When evaluated on the KTM dataset the deep features used as the input to each algorithm were those from layers  $conv1\text{-}1$ ,  $conv3\text{-}4$  and  $conv4\text{-}2$  of Model-D. BBS, CoTM and QATM have been tested on BBS data by their authors using different deep features, so we also compare our results to these earlier published results.

The comparison results are summarised in Table 4. All methods expect QATM and BBS produce improved results using the proposed deep features than when using color features. This is true for both datasets. Of the methods that have previously been applied to deep features the performances of two (NCC and QATM) are improved, and that of two others (BBS and CoTM) are made worse by using our method of defining the deep feature-space.

**Table 4** Quantitative comparison of the performance of different template matching algorithms using different input features

Methods	Feature				
	BBS dataset			KTM dataset	
	Color	Deep	Deep (Proposed)	Color	Deep (Proposed)
	AUC				
SSD	0.46	–	0.54	0.42	0.54
NCC	0.48	0.63 [13]	0.67	0.42	0.67
ZNCC	0.54	–	0.67	0.48	0.67
BBS	0.55	0.60 [14]	0.54	0.44	0.55
CoTM	0.54 <sup>a</sup>	0.67 [12]	0.64	0.51	0.56
DDIS	0.64	–	0.66	<b>0.63</b>	0.68
QATM	–	0.62 <sup>b</sup>	0.66	–	0.64
DIM	<b>0.69</b>	–	<b>0.73</b>	0.60	<b>0.71</b>

<sup>a</sup> We were unable to reproduce the result 0.62 reported in the paper [12] using code supplied by the authors of CoTM, our different result is shown in the table

<sup>b</sup> The authors of QATM report an AUC score of 0.69 when this method is applied to the BBS dataset [11]. However, examining their source code we note that this result is produced by setting the size of the predicted bounding box equal in size to the width and height of the ground truth bounding box. Other methods are evaluated by setting the size of the predicted bounding box equal to the size of the template (i.e. without using knowledge of the ground truth that the algorithm is attempting to predict). We have re-tested QATM using the standard evaluation protocol and our result for the original version of QATM is 0.62. As QATM is designed to work specifically with a CNN it was not applied directly to color images

However, it should be noted that simple metrics for comparing image patches such as NCC and ZNCC produce near state-of-the-art performance when applied to our proposed deep feature-space, outperforming much more complex methods of template matching such as BBS, CoTM, and QATM. One known weakness of BBS is that it may fail when the template is very small compared to target image [14]. This may explain the particularly poor results of this method when applied to the KTM dataset.

DIM achieves the best results on both datasets when applied to deep features. DIM performs particularly well on the BBS dataset producing an AUC of 0.73 which, as far as we are aware, makes it the only method to have scored more than 0.7 on this dataset. The DIM algorithm also produces state-of-the-art performance on the KTM dataset when applied to deep features. When applied to color features, the results are good, although not as good as DDIS on the KTM dataset. This is because small templates in the KTM dataset sometimes contain insufficient detail for the DIM algorithm to successfully distinguish the object. Using deep features enhances the discriminability of small templates sufficiently that the performance of DIM increases significantly. The results demonstrate that the proposed approach

is effective at extracting distinguishable features which lead to robust and accurate template matching.

## 5 Conclusions

Our results demonstrate that slightly increasing the shape bias of a CNN (by changing the method of training the network) produces more distinguishable features in which template matching can be achieved with greater accuracy. By running a large number of experiments we determined the best combination of convolutional features from our shape-biased VGG19 on which to perform template matching with the DIM algorithm. This same feature-space was shown to improve the performance of most other template matching algorithms as well. The DIM algorithm applied to our new feature-space produces state-of-art results on two benchmark datasets.

**Acknowledgements** The authors acknowledge use of the research computing facility at King's College London, Rosalind (<https://rosalind.kcl.ac.uk>), and the Joint Academic Data science Endeavour (JADE) facility. This research was funded by China Scholarship Council.

## References

1. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fullyconvolutional siamese networks for object tracking. In: European Conference on Computer Vision. pp. 850–865. Springer, Berlin (2016)
2. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Robust visual tracking via hierarchical convolutional features. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(11), 2709–2723 (2018)
3. Ahuja, K., Tuli, P.: Object recognition by template matching using correlations and phase angle method. *Int. J. Adv. Res. Comput. Commun. Eng.* **2**(3), 1368–1373 (2013)
4. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
5. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **47**(1–3), 7–42 (2002)
6. Chhatkuli, A., Pizarro, D., Bartoli, A.: Stable template-based isometric 3d reconstruction in all imaging conditions by linear least-squares. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 708–715 (2014)
7. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.: PCANet: a simple deep learning baseline for image classification? *IEEE Trans. Image Process.* **24**(12), 5017–5032 (2015)
8. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
9. Liang, M., Hu, X.: Recurrent convolutional neural network for object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3367–3375 (2015)
10. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3109–3118 (2015)

11. Cheng, J., Wu, Y., AbdAlmageed, W., Natarajan, P.: QATM: quality-aware template matching for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11553–11562 (2019)
12. Kat, R., Jevnisek, R., Avidan, S.: Matching pixels using co-occurrence statistics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1751–1759 (2018)
13. Kim, J., Kim, J., Choi, S., Hasan, M.A., Kim, C.: Robust template matching using scale-adaptive deep convolutional features. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 708–711. IEEE (2017)
14. Oron, S., Dekel, T., Xue, T., Freeman, W.T., Avidan, S.: Best-buddies similarity—robust template matching using mutual nearest neighbors. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(8), 1799–1813 (2017)
15. Talmi, I., Mechrez, R., Zelnik-Manor, L.: Template matching with deformable diversity similarity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 175–183 (2017)
16. Kriegeskorte, N.: Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015)
17. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. [arXiv:1811.12231](https://arxiv.org/abs/1811.12231) (2018)
18. Spratling, M.W.: Explaining away results in accurate and tolerant template matching. *Pattern Recogn.* 107337 (2020)
19. Kersten, D., Mamassian, P., Yuille, A.: Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004)
20. Spratling, M.W.: Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Comput.* **24**(1), 60–103 (2012)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
23. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
24. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1834–1848 (2015)
25. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: algorithms and benchmark. *IEEE Trans. Image Process.* **24**(12), 5630–5644 (2015)

# Reach on Visual Image Restoration Method for AUV Autonomous Operation: A Survey



Teng Xue<sup>✉</sup>, Jing Zhang<sup>✉</sup>, and Tianchi Zhang<sup>✉</sup>

**Abstract** AUV (Autonomous Underwater Vehicle) is playing an increasingly important role in the ocean development process. Autonomous operation capability is the development direction of AUV, which mainly focuses on underwater observation, marine resource development, sampling and inspection, and autonomous reconnaissance. People pay more attention to military, political, economic and social needs. The autonomous operation of AUV has the advantages of a wide range of activities and good concealment. AUV can obtain images through an optical vision system. The image is prone to distortion, low saturation, blur, etc. The restoration of underwater images has been attracting more and more research work. Underwater light scattering and absorption are the root reasons of image quality degradation. This paper introduces the latest development of underwater image restoration methods, analyzes and summarizes the research status of several types of methods. The latest data set is given, several image evaluation indicators are summarized, and the future development direction is pointed out.

**Keywords** Underwater image restoration · Autonomous underwater vehicle · Deep learning

## 1 Introduction

AUV is a new generation of underwater robots. It has a wide range of motion, good mobility, safety and intelligence [1]. AUV becomes an important tool for completing kinds of underwater tasks. AUV can be used in the civil field for pipeline laying,

---

T. Xue · J. Zhang

School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

T. Zhang (✉)

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China

e-mail: [zhangtianchi@cqjtu.edu.cn](mailto:zhangtianchi@cqjtu.edu.cn)

seabed survey, data collection, seabed construction, and maintenance of equipment. AUV can also be used in the military field for reconnaissance, mine-laying, and rescue activities [2]. Because the underwater robot without cable has the advantages of no cable limitation and good concealment, all parties have begun to have an interest in AUV [2]. Due to AUV's mobility and low noise, AUV plays an important and irreplaceable role in marine development and coastal defense [3, 4].

Underwater imaging consists of direct signal and background scattering signal. Because light scattering and absorption affect light propagation, light propagation in water is attenuated exponentially [5]. In addition, the light propagation of different wavelengths is different. For wavelengths of different colors, about ten meters deep, the red wavelength disappears, at 15 m deep, the orange wavelength disappears, and at 20 m, the yellow wavelength disappears, and then the green and purple wavelengths disappear at a depth of about 35 m. The underwater image is mainly blue because the blue wavelength moves in the water for the longest time [6]. The small particles dissolved in water and the inherent noise of underwater imaging system can also lead to the distortion of underwater image [5]. In addition, different waters have different light conditions, so the degree of underwater degradation is not the same. Considering the objective factors, such as the instability of AUV optical vision system camera, poor focusing, the mobility of underwater animals and the noise of various sensors and so on, the underwater image will be degraded to varying degrees.

The work of this paper mainly has the following four points:

1. Four kinds of underwater image restoration methods are summarized.
2. The existing data sets for underwater image restoration are introduced.
3. Several evaluation indicators are summarized.
4. Made a conclusion.

## 2 Method of Underwater Image Restoration

This paper briefly summarizes the existing methods and divides them into the following four categories: based on dedicated hardware driver; based on prior information; based on deep network learning; based on Jaffe-McGlamery model.

### 2.1 Special Hardware Driver Methods

Xie et al. [7] proposed a related algorithm for underwater degraded image restoration and established a corresponding model to prove this algorithm. On this basis, denoising processing of underwater degraded images using Gaussian kernel function. According to the denoising results, the underwater polarization imaging model was established. Wang [8] proposed an image restoration algorithm for laser underwater optical imaging. In terms of the characteristics of noise and underwater image classification, Wiener filtering and constrained minimum power filtering are shown.

Ishibashi [9] proposed an underwater camera model and described its outline. Experimental results show that this method has high practicability and good effects in stereo imaging. Nascimento et al. [10] proposed a fully automatic underwater image restoration method based on the physical model of light propagating in the medium. This method can obtain images from different angles. These images are images in the same environment. The core of the method is an iterative algorithm based on contrast measurement. It can automatically estimate all the parameters of the model with a small amount of calculation and has a high accuracy rate [10]. Chen et al. [11] established an imaging model on the basis of beam transmission, which is intended to be applied to image super-resolution reconstruction tasks within underwater imaging system of range-gated pulsed laser. We should know that although the hardware solution has certain effectiveness, it is not suitable for dynamic acquisition. Therefore, most of the underwater image restoration uses the other three methods.

## 2.2 Prior Image Information Methods

He et al. [12] put forward the dark channel prior method for the first time. He observed the difference between the normal image and the haze image and took a big step towards the image restoration. We know that in the outdoor fog-free image, in most non-sky areas, each color channel will have some lower pixel values, that is, dark pixels. However, in haze, because of the water in the air and the scattered light of dust particles, the whole picture is gray and white, so the dark channel is no longer black. On the contrary, the denser the fog area is, the closer the pixels of the dark channel are to the appearance of the fog itself caused by air scattering. We can know that underwater image and haze image have many similarities, such as blur, color distortion, low color saturation and so on.

Physical model:

$$I = Jt + A(1 - t) \quad (1)$$

In (1), where  $I$  is the haze image,  $J$  is the clear image,  $A$  is the atmospheric scattered light, and  $t$  is the medium transmittance. Similarly, with the introduction of underwater image, I refer to the image obtained by the imaging equipment,  $J$  is the original undistorted image,  $A$  is the underwater global background light, and  $t$  is the transmittance of the underwater channel.

$$t = e^{-\beta d} \quad (2)$$

In (2), the transmittance of the underwater channel is determined by the attenuation coefficient  $\beta$  and the depth of field  $d$ . Based on the algorithm of image defogging, a variety of underwater image restoration methods are proposed.

Therefore, many scholars are based on the a priori idea of the dark channel, oceanographer Derya Akkaynak and engineer Tali Treibitz [13] to correct the

degraded color of the image, a “sea thru” algorithm was created, eliminate color cast and scattering of light, making the photo as clear as if taken in a place without water, bringing higher clarity and realism to underwater photos. Galdran et al. [14] proposed a short-wavelength color restoration method for underwater images, which is based on the red channel, thereby restoring the lost contrast. Experimental results show that this method performs better than most existing methods, this technology can effectively handle artificial lighting areas, and can achieve natural color correction and visibility improvement. Hanmant and Ingle [15] proposed a method for estimating the depth of underwater scenes for image restoration, which is based on image blur and light absorption and enhancement in image formation models (IFM). In the past, IFM-based image restoration methods for estimating scene depth were based on dark channel priors or based on maximum intensity priors. The lighting conditions in underwater images often invalidate these images, resulting in poor restoration effects. This method can relatively accurately estimate the depth of the scene in the underwater image. Zhang and Peng [16] proposed a new underwater imaging model, based on the characteristics of underwater imaging and underwater illumination and put forward a joint prior-based estimation method of underwater image medium transmission to predict underwater images Medium transmission. In addition, the global background light commonly used in the underwater image restoration method can be replaced by the color of the light source to correct the color of the underwater degraded image [17]. However, there is a difference between the spread of light in the ocean and the propagation in the atmosphere, which makes the model imperfect when applied to underwater operations. Based on this situation, Derya and Tali [17] made improvements and proposed an improved imaging model, pointing out that the errors generated by the currently commonly used underwater imaging models have not been considered. The improved model can motivate people to continue to develop better models for correcting complex underwater images.

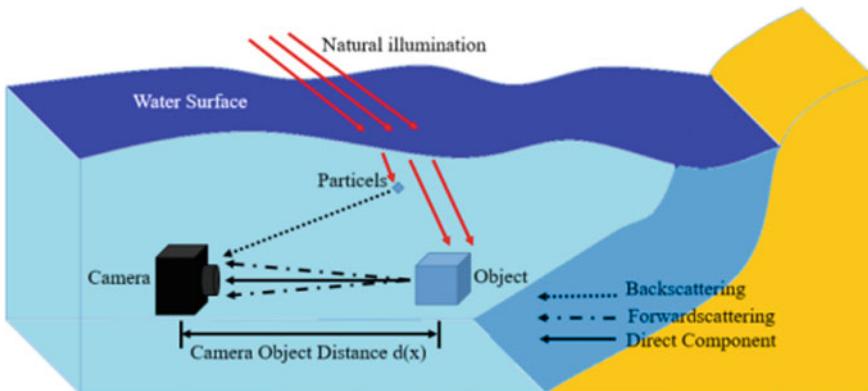
### 2.3 Deep Network Learning Method

Keyan et al. [18] had proposed a convolutional neural network (CNN) that can be effectively used for underwater image restoration. Yang et al. [19] put forward a deep method for underwater image enhancement. Yu et al. [20] proposed underwater GAN, which is an underwater image restoration method based on a conditional generation confrontation network. Nan et al. [21] proposed UWGAN for color restoration and noise removal of real underwater images. Chongyi et al. [22] proposed a weakly supervised method to correct color distortions and thus perform color transfer. This method reduces the training requirements for underwater images and allows underwater images to be taken in unknown locations. Experimental results show that this method is superior to existing methods due to its good resilience. Dudhane et al. [23] proposed an underwater image recovery generator based on a dense residual network of channels. This dense residual network is composed of two parts, which are used to extract colors and other related features, and to remove underwater image blur.

Using deep learning methods to restore underwater images takes a long time to train, but if a reasonable network structure is designed, it can have a strong learning ability and can effectively restore underwater images. Therefore, recovery methods based on deep networks have developed rapidly and have good prospects.

## 2.4 Jaffe-McGlamery Model-Based Methods

First, we need to understand the underwater imaging model by Jaffe and McGlamery [24, 25]. Underwater images can be expressed as direct components, forward scattered components and backscattered components. Among them, the forward scattered light is reflected by the target surface or suspended particles in the water and enters the imaging system, thereby blurring the image. After natural light enters the water, the backscattered light is affected by the scattering of suspended particles, so that the light entering the imaging system will cause the image obtained by the imaging system to display low contrast. Figure 1 shows the underwater imaging model. As shown in Fig. 1, Emanuele Trucco and Adriana T. Olmos-Antillon based the Jaffe–McGlamery model proposed a self-tuning image restoration filter based on simplified model. The algorithm performs self-correction and estimates the best parameter value for each image [26]. Cheng et al. [27] simplified the Jaffe-McGlamery model and defined a red channel before estimating the background light and transmittance. On this basis, in order to eliminate the ambiguity of the underwater degradation model, a simple and effective low-pass filter is designed. This algorithm is eliminating scattering While absorbing, it effectively restores the underwater image.



**Fig. 1** The underwater image formation model

### 3 Data Set

Dalian University of Technology Geometric Calculation and Intelligent Media Technology Research Team proposed a data set called RUIE, which is a Real Underwater Image Enhancement for testing underwater images and recovery algorithm in 2020, it has a lot of data and the degree of light scattering effect. The underwater biological label image can be applied to the training and testing of underwater target detection algorithms. The RUIE dataset is the first large underwater real image database that is specifically designed for multi-angle algorithm.

In a paper published in 2017, Jian et al. [28] introduced a large underwater image dataset, called detection-OUC-VISION underwater image dataset, which consists of 4400 real underwater images and 220 single images of underwater targets. The database can be applied to underwater salient target detection or saliency detection and will contribute to the development of technology in this field in the future.

Chongyi et al. [29] constructed a real underwater image enhancement benchmark (UIEB), this data set is composed of 950 real underwater images, and the scale is large. These images were taken under different light sources, and on this basis, the most advanced single-frame underwater enhancement and restoration methods at home and abroad have been comprehensively studied.

Song et al. [30] developed artificially labeled background light (MABL) database in 2020, containing 500 underwater images. This is the first database currently used for background light evaluation of underwater images.

### 4 Evaluating Indicator

The evaluation indicator can evaluate the quality of the image restoration result very well. Classic objective evaluation indicators include Ambiguity, Clarity, Peak Signal to Noise Ratio, Structural Similarity and so on.

Subjective evaluation method: The subjective evaluation of observers is the most common and direct method of image quality evaluation. Chongyi et al. [22] invited 10 members to rate the results. These ten people are experienced in image processing. Each member does not know which result the method proposed by the author is, and there is no time limit in the process of scoring by members. Therefore, a relatively fair average visual quality score can be obtained.

### 5 Conclusion

This paper introduces several methods of AUV's autonomous operation of visual image restoration and summarizes the common problems in these methods, which can help researchers gain some knowledge about underwater image restoration. This

paper takes four different methods as a starting point and introduces four different types of image restoration methods, the following work can be carried out based on these methods. It should be noted that underwater image restoration technology is constantly evolving and needs continuous improvement.

**Acknowledgements** This research is supported by: (1) 2020–2022 National Natural Science Foundation of China under Grand (Youth) No. 52001039. (2) 2020–2022 Funding of Shandong Natural Science Foundation in China No. ZR2019LZH005.

## References

1. Haiyan, L., Hao, W.: Application: research status, prospect of machine vision technology in underwater robot. *Inf. Recording Mater.* **20**(09), 18–19 (2019)
2. Research status of cable-free autonomous underwater vehicle (AUV). <https://wenku.baidu.com/view/2e7e2ebac77da26925c5b0ba.html>. Last accessed 2012/02/25
3. Tanakitkorn, K., Wilson, P.A., Turnock, S.R., Phillips, A.B.: Depth control for an over-actuated, hover-capable autonomous underwater vehicle with experimental verification. *Mechatronics* **41**, 67–81 (2017)
4. Maria, L.C., Giuseppe, O.: A robust observer-based fault tolerant control scheme for underwater vehicles. *J. Dyn. Syst. Meas. Control* **136**(3), 1–11 (2014)
5. Jichang, G., Chongyi, L., Chunle, G., Chen, S.: Research progress of underwater image enhancement and restoration methods. *Chin. J. Image Graph.* **22**(03), 273–287 (2017)
6. Abril, L., et al.: Color correction of underwater images for aquatic robot inspection. In: Rangarajan, B.C., Vemuri, A.L., Yuille (eds.) *Lecture Notes in Computer Science*, vol. 3757, pp. 60–73. Springer, Berlin (2005)
7. Xie, Q., Yang, W., Lu, Z.: An underwater degraded image restoration algorithm based on polarization imaging. *Computer Simulation* **37**(12), 249–252+257 (2020)
8. Wang, H.: Research on image restoration clarification algorithm of laser underwater optical imaging. *Laser J.* **37**(9), 4 (2016)
9. Ishibashi, S.: The study of the underwater camera model. *Oceans. IEEE*, pp. 1–6 (2011)
10. Nascimento, E.R., Campos, M.F.M., Barros, W.F.: Stereo based structure recovery of underwater scenes from automatically restored images. In: *Proceedings of the 22th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 330–337 (2009)
11. Chen, Y., Yang, B., Xia, M., et al.: Model-based super-resolution reconstruction techniques for underwater imaging. *Proc. SPIE* **8332**, 1 (2012)
12. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 2341–2353 (2011)
13. Treibitz, D.A.T.: A Revised Underwater Image Formation Model. *IEEE* (2018)
14. Galdran, A., Pardo, D., Picón, A., et al.: Automatic red-channel underwater image restoration. *J. Vis. Commun. Image Represent.* **26**, 132–145 (2015)
15. Hammante, B.P., Ingle, M.: Underwater image restoration based on light absorption. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE (2018)
16. Zhang, M., Peng, J.: Underwater image restoration based on a new underwater image formation model. *IEEE Access* **6**, 58634–58644 (2018)
17. Derya, A., Tali, T.: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6723–6732 (2018)
18. Keyan, W., Yan, H., Jun, C., Xianyun, W., Xi, Z., Yunsong, L.: Underwater image restoration based on a parallel convolutional neural network. *Remote Sens.* **11**(13), 1591 (2019)

19. Yang, W., Jing, Z., Yang, C., et al.: A deep CNN method for underwater image enhancement. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE (2017)
20. Yu, X., Qu, Y., Ming, H.: Underwater-GAN: underwater image restoration via conditional generative adversarial network. In: International Conference on Pattern Recognition. Springer, Cham (2018)
21. Nan, W., Yabin, Z., Fenglei, H., Haitao, Z., Jingzheng Y.: UWGAN: underwater GAN for real-world underwater color restoration and dehazing (2019)
22. Chongyi, L., Jichang, G., Chunle, G.: Emerging from water: underwater image color correction based on weakly supervised color transfer. *IEEE Signal Process. Lett.* **25**(3), 323–327 (2018)
23. Dudhane, A., Hambarde, P., Patil, P., Murala, S.: Deep underwater image restoration and beyond. *IEEE Signal Process. Lett.* **27**, 675–679 (2020)
24. McGlamery, B.L.: A computer model for underwater camera systems. In: SPIE Ocean Optics, vol. 208, pp. 221–231 (1979)
25. Jaffe, J.S.: Computer modeling and the design of optimal underwater imaging systems. *IEEE J. Oceanic Eng.* **15**(2), 101–111 (1990)
26. Trucco, E., Olmos-Antillon, A.T.: Self-tuning underwater image restoration. *IEEE J. Oceanic Eng.* **31**(2), 511–519 (2006)
27. Cheng, C., Sung, C., Chang, H.: Underwater image restoration by red-dark channel prior and point spread function deconvolution. In: 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 110–115 (2015)
28. Jian, M., Qi, Q., Dong, J., et al.: The OUC-vision large-scale underwater image database. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 1297–1302 (2017)
29. Chongyi, L., et al.: An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* **29**, 4376–4389 (2020)
30. Song, W., Wang, Y., Huang, D., et al.: Enhancement of underwater images with statistical model of background light and optimization of transmission map. *IEEE Trans. Broadcast.* **66**(1), 153–169 (2020)

# On Improving Perceptual Image Hashing Using Reference Image Construction



Xinran Li  and Zichi Wang 

**Abstract** This paper proposes an improved framework for perceptual image hashing, which is able to improve existing image hashing schemes. In our framework, hash sequence contains two parts. The first one is generated by an existing perceptual image hashing scheme on an input image. Then, a reference image is constructed by executing Wiener filtering on the input image. Subsequently, the same hashing scheme is executed on the reference image to obtain the second part. Finally, the two parts are combined together to form the improved hash. With the improved hash, the properties of main content of image can be fully extracted. Experimental results show that the performances of existing perceptual image hashing schemes, e.g., perceptual robustness, discrimination capability, can be improved using our framework.

**Keywords** Digital image · Perceptual hashing · Improvement

## 1 Introduction

The technique of perceptual image hashing aims to map a given image into a short sequence (called image hashing) with a fixed length [1]. The generated hash sequence is robust to conventional content-preserving manipulations, e.g., filtering, JPEG compression, geometrical transform and noising, and is significantly different for perceptually distinct images. These properties cannot be achieved by typical cryptographic hash functions which are very sensitive to tiny modifications on input data [2]. For this reason, perceptual image hashing has been well developed in the past two decades. There are a lot of important applications for perceptual image hashing, such as image authentication [3], image retrieval [4], digital forensics [5].

---

X. Li (✉)

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

e-mail: [191550054@st.usst.edu.cn](mailto:191550054@st.usst.edu.cn)

Z. Wang

School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

The derivation of perceptual image hashing can be traced back to [6], in which the concept of image hashing was proposed. After that, many effective perceptual image hashing schemes had been developed [7]. Traditional proposals were based on manual hash extraction, mainly included three stages: pre-processing, feature extraction and hash generation. Many techniques were employed for perceptual image hashing, such as DCT (Discrete Cosine Transform) [8], SVD (Singular Value Decomposition) [9], and CS (Compressed Sensing) [10]. Besides, some schemes were based on ring partition [11–13]. In [11], the input image was divided into different rings after normalization. Then, the ring-based entropies were employed to obtain the final hash sequence. In [12], a rotation-invariant secondary image was constructed to give image hash the ability to resist rotation. In [13], the rotation robustness was further enhanced by ring partition and invariant vector distance. Recently, deep learning based perceptual image hashing schemes achieved good performance, in which the stages of feature extraction and hash generation were combined by deep networks automatically [14–16]. In [14], deep convolutional neural network was used to learn binary image hashing. A hash layer was added into the deep network to learn the hash function with visual feature. In [15], deep network was trained in a layer-wise manner to progressively improve robustness. The authors proved that the hierarchical architecture is beneficial to the security of hash function. In [16], the training set structure was adjusted dynamically with multiple constraints to guarantee the robustness and discrimination simultaneously.

Although more and more perceptual image hashing schemes were designed and achieved good performance, the cost for developing a new hashing scheme is high, and the utilization of existing schemes is insufficient. In addition, the generality is poor if it only improves a certain scheme [17]. Therefore, a general framework which is able to improve existing perceptual image hashing schemes is desirable. Although the existing hashing schemes are effective for digital images, they still can be further improved.

Robustness to content-preserving manipulations is an important property of perceptual image hashing. This paper proposes a universal framework to improve perceptual image hashing with the help of content-preserving manipulation. The improved hash contains two parts. An existing image hashing scheme is executed on an input image to obtain the first part. With a content-preserving manipulation, a reference image is constructed from the input image. As a result, main content of the input image is filtrated. The second part is obtained by employing the same hashing scheme on the reference image. Finally, the two parts are combined to form the new hash. The novelty and contributions of this paper are summarized as follows:

1. We propose a general hash improved framework for digital images, which is able to improve existing image hashing schemes. In our framework, we propose the strategy of reference image construction which is effective for perceptual image hashing.
2. Using our framework, the performances (such as perceptual robustness and discrimination capability) of existing image hashing schemes (including manual

and deep learning based schemes) can be improved without increasing the length of hash.

## 2 Proposed Framework

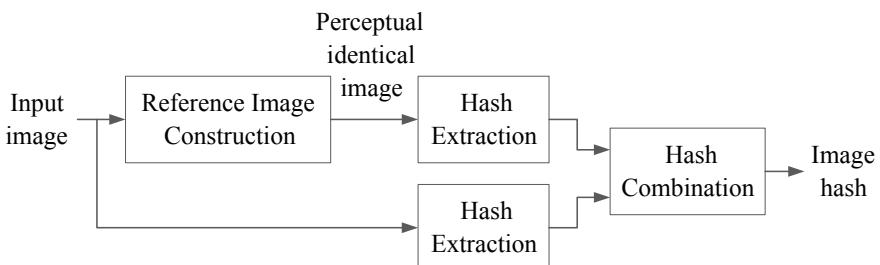
In this paper, we propose a general improved framework for perceptual image hashing using the strategy of reference image construction. As shown in Fig. 1, an existing image hashing scheme is employed on an input image firstly. Then, a reference image is constructed from the input image using a content-preserving manipulation to filtrate the main content of the input image more efficiently. The obtained reference image is identical to the input image semantically. After that, the same hashing scheme is employed again on the reference image. The obtained two hashes are combined to form the final hash sequence. The details are as follows.

### 2.1 Reference Image Construction

To filtrate the main image content more efficiently, a reference image is constructed. To achieve satisfactory perceptual robustness, the constructed reference image should be semantically identical with the input image. Wiener filter is a superior tool for reference construction with minimal MSE (Mean Squared Error) [18, 19], which reserves the main content of an image without bringing in additional noise (some content-preserving manipulations will bring in unnecessary noise, such as speckle noising, JPEG compression). For this reason, we use Wiener filter to construct a reference image which is close to the input one. With proper parameters, the obtained reference image is able to supplement necessary information about the input image.

For an input image  $\mathbf{X}_o$  sized  $M \times N$ , a reference image  $\mathbf{X}_r$  is obtained after Wiener filtering. The value  $m$  of MSE between  $\mathbf{X}_o$  and  $\mathbf{X}_r$  is

$$m = E\{[x(n) - x_r(n)]^2\}, \quad (1)$$



**Fig. 1** Architecture of the proposed framework

where the operator  $E\{\cdot\}$  is mathematical expectation, and  $n \in \{1, 2, \dots, MN\}$ . To achieve a Wiener filter, the indicators mean and standard deviation are estimated using a specified local neighbourhood of each pixel. Specifically,  $\mathbf{X}_o$  is filtered by pixel-wise adaptive Wiener filtering. As shown in Eqs. (2) and (3), the mean  $\mu$  and standard deviation  $\sigma$  can be calculated using neighbourhoods with a given size  $s$  (will be given in experiments).

$$\mu = \frac{1}{s^2} \sum_{i,j \in C} x_o(i, j), \quad (2)$$

$$\sigma = \sqrt{\frac{1}{s^2} \sum_{i,j \in C} x_o^2(i, j) - \mu^2}, \quad (3)$$

where  $x(i, j)$  is the  $(i, j)$ -th pixel of image  $\mathbf{X}_o$ , and  $C$  is the  $s \times s$  neighbourhood of  $x(i, j)$ . The power of the difference between  $\mathbf{X}_o$  and  $\mathbf{X}_r$  can also be estimated using the statistical results calculate from  $\mathbf{X}_o$ . Finally, reference image can be constructed using Eq. (4),

$$\mathbf{X}_r = \mathbf{X}_o \otimes \mathbf{F}, \quad (4)$$

where  $\mathbf{F}$  is a 2D Wiener filter, and “ $\otimes$ ” is the operation of convolution. The filtered version is perceptual identical with the input image.

The reference image is a supplement of the input image, which is helpful to capture the main content of the input image. This is significant to perceptual image hashing since main content of an image is important to perceptual robustness.

## 2.2 Hash Generation

With the input image  $\mathbf{X}_o$  and reference image  $\mathbf{X}_r$ , an existing image hashing scheme is employed to extract hashes respectively. The method employed for hash extraction will be discussed in experiments, in which three state-of-the-art image hashing methods RE [11], RP-NMF [12], and DAE-NN [15] are employed for comparison.

Denote the hashes of  $\mathbf{X}_o$  and  $\mathbf{X}_r$  as  $\mathbf{H}_o = \{h_o(i)\}$  and  $\mathbf{H}_r = \{h_r(i)\}$ ,  $i \in \{1, 2, \dots, k\}$ , where  $k$  is the length of hash sequence. Another element to obtain the final hash sequence is the way to combine  $\mathbf{H}_o$  and  $\mathbf{H}_r$ . Since the lengths of  $\mathbf{H}_o$  and  $\mathbf{H}_r$  are the same, we empirically achieve the combination by the straightforward operation of averaging. That means the final hash sequence  $\mathbf{H} = \{h(i)\}$  is the average result of  $\mathbf{H}_o$  and  $\mathbf{H}_r$ , as shown in Eq. (5).

$$h(i) = \frac{h_o(i) + h_r(i)}{2}. \quad (5)$$

In this way, the performances of existing image hashing schemes can be improved without increasing the length of hash. During the applications for perceptual image hashing such as image authentication, the length of hash sequence is highly related to computational complexity.

### 3 Experimental Results

To check the effectiveness of our framework, we conduct some groups of experiments. Firstly, we set up the experimental environments. Then, we discuss the perceptual robustness and discrimination capability. Finally, we provide the results of image authentication.

#### 3.1 Experiment Setup

We employ the image dataset UCID [20] for experiments, which contains 1338 uncompressed color images sized  $512 \times 384$ . Moreover, some popular test images (shown in Fig. 2), which are widely utilized in the fields of image processing, are also employed for experiments.

To measure the distance  $D(\mathbf{H}_1, \mathbf{H}_2)$  of two hashes  $\mathbf{H}_1 = \{h_1(i)\}$  and  $\mathbf{H}_2 = \{h_2(i)\}$ ,  $i \in \{1, 2, \dots, k\}$ , MSE is employed in this section, as shown in Eq. (6). Smaller value of  $D(\mathbf{H}_1, \mathbf{H}_2)$  means higher similarity between the two hashes. Usually, two images can be judged as perceptually identical if the distance  $D$  between the corresponding two hashes is smaller than a given threshold  $T$ . Otherwise, the two images are perceptually distinct.

$$D(\mathbf{H}_1, \mathbf{H}_2) = \frac{1}{k} \sum_{i=1}^k [h_1(i) - h_2(i)]^2. \quad (6)$$

For comparison, three popular image hashing schemes RE [11], RP-NMF [12], and DAE-NN [15] are employed. The filter size  $s$  in Subsection 2. A for the three schemes are determined as 3, 5, 6 by experiments, respectively. For convenience,



**Fig. 2** Several popular test images. **a** Lena; **b** aerial; **c** peppers; **d** earth; **e** car

**Table 1** Parameter settings of content-preserving manipulations

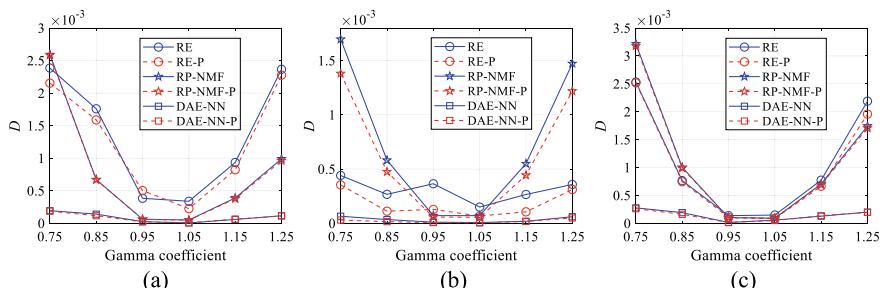
Manipulation	Parameter	Values
Gamma correction	Gamma coefficient	0.75, 0.85, 0.95, 1.05, 1.15, 1.25
Speckle noise	Variance	0.05, 0.06, 0.07, 0.08, 0.09, 0.1
Wiener filtering	Window size	3, 5, 7, 9, 11, 13

the hash values of RE and RP-NMF are restricted to 0 to 1 in our experiments. To get perceptually identical images, three content-preserving manipulations Gamma Correction, Speckle Noise, and Wiener Filtering are used in our experiments. The corresponding parameter settings are listed in Table 1. In total, 18 perceptual images will be produced for each input image.

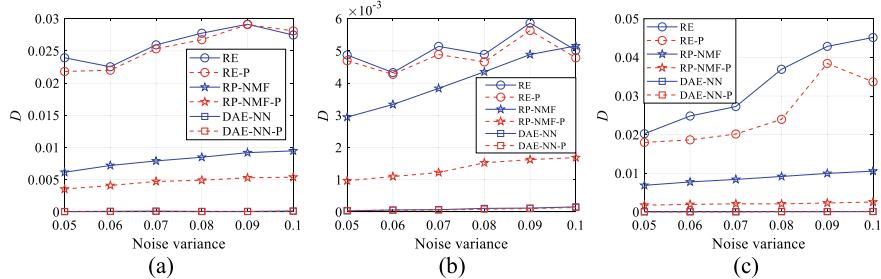
### 3.2 Perceptual Robustness

Robust to content-preserving manipulation is an important property of perceptual image hashing. It means that the distances between hash pairs of two perceptually identical images should be small enough. The comparisons of perceptual robustness between our framework and the schemes RE, RP-NMF, DAE-NN tested on images Lena, Aerial, and Peppers are shown in Figs. 3, 4 and 5. The ordinate stands for the distance between hash pair of an input image and its perceptually identical version, while the abscissa stands for the value of parameter in Table 1. Where “RE-P”, “RP-NMF-P”, and “DAE-NN-P” stand for the improved versions using our framework.

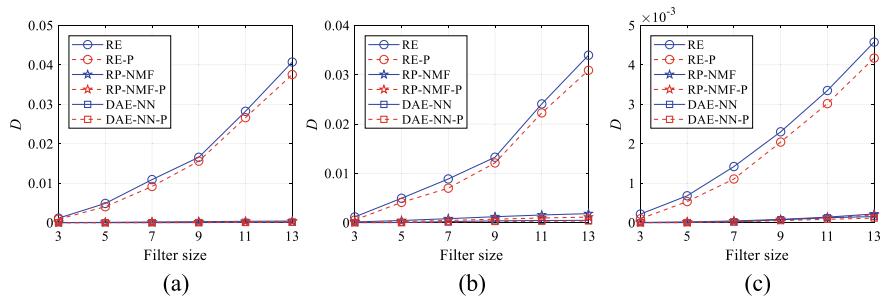
It can be seen from Figs. 3, 4 and 5 that the hash distance of existing hash schemes tested on perceptually identical images are shortened in most cases after our framework is used. That means the perceptual robustness of existing hash methods is improved by using our framework. To further verify the effectiveness of our framework, all the 1338 images in UCID are used as input images. The average distance comparisons of image hashing schemes tested on these images are shown in Fig. 6.



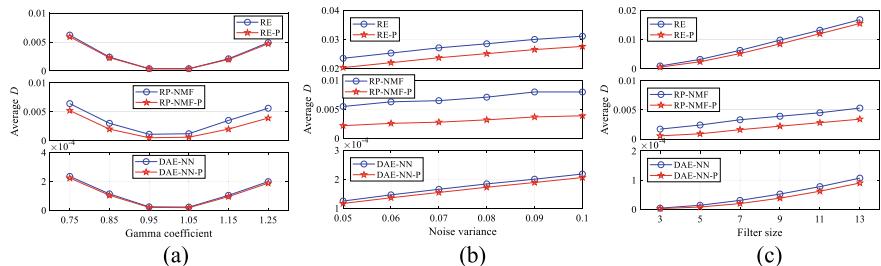
**Fig. 3** Hash distances comparisons of gamma correction on **a** Lena; **b** aerial; **c** peppers



**Fig. 4** Hash distances comparisons of speckle noise on **a** Lena; **b** aerial; **c** peppers



**Fig. 5** Hash distances comparisons of WIENER FILTERING on **a** Lena; **b** aerial; **c** peppers



**Fig. 6** Hash distances comparisons on UCID for **a** Gamma correction; **b** speckle noise; **c** wiener filtering

The results also indicate that our framework is effective in improved the perceptual robustness of existing hash schemes. Specifically, using our framework, the average improvement on MSE of RE for Gamma Correction is 0.77%, while 2.04% for Speckle Noise, and 2.02% for Wiener Filtering. For RP-NMF, the average improvement is 5.29%, while 9.26% for Speckle Noise, and 5.58% for Wiener Filtering. For DAE-NN, the corresponding improvements are 1.16%, 1.03%, and 3.78%, respectively.

**Table 2** Average hash distances comparison

Images	RE/RE-P	RP-NMF/RP-NMF-P	DAE-NN/DAE-NN-P
300 images in UCID	0.14882 <b>0.14884</b>	0.046806 <b>0.046961</b>	0.019591 <b>0.019600</b>

The bold stand for the better performances

**Table 3** Hash distances between perceptually distinct images

Schemes	Lena and aerial	Lena and peppers	Lena and earth	Lena and car
RE	0.13686	0.20290	0.10636	0.04201
RE-P	<b>0.13734</b>	<b>0.20294</b>	<b>0.10805</b>	<b>0.04353</b>
RP-NMF	0.03912	0.00944	0.00378	0.03054
RP-NMF-P	<b>0.03971</b>	<b>0.01022</b>	<b>0.00431</b>	<b>0.03283</b>
DAE-NN	0.02100	0.02327	0.01300	0.02251
DAE-NN-P	<b>0.02145</b>	<b>0.02342</b>	<b>0.01304</b>	<b>0.02254</b>

The bold stand for the better performances

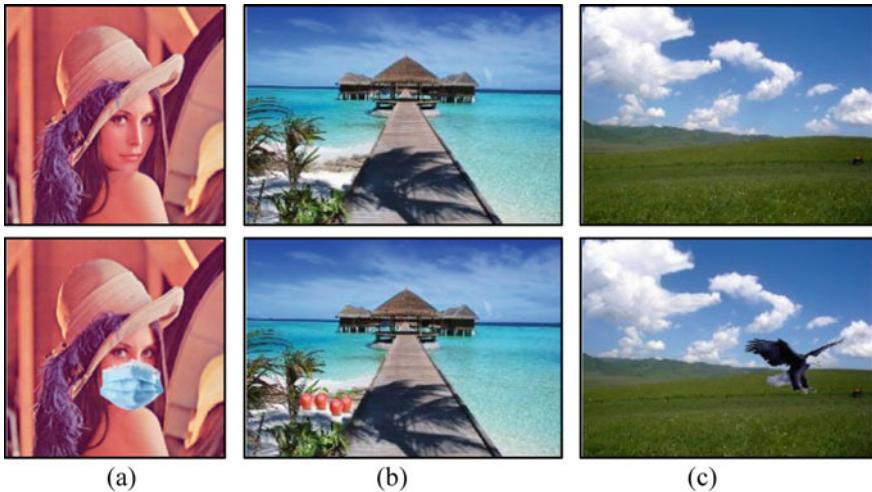
### 3.3 Discrimination Capability

Contrary to perceptual robustness, for perceptually distinct images, the hash distance should be great enough. To test the discrimination capability, we firstly extract the hashes of 300 images selected from UCID. Then, for each hash, we calculated the average distance  $D$  between it and all the other 299 hashes. Totally, there are  $(300 \times 299)/2 = 44,850$  distances obtained for all the hash pairs of the 300 perceptually distinct images. Table 2 lists the average values of the 44,850 distances for different image hashing schemes. Meanwhile, some distances calculated from the images in Fig. 2 are listed in Table 3.

It is clear that the hash distance of existing hash schemes tested on perceptually distinct images is enlarged after our framework is used. Therefore, the discrimination capability of existing image hashing schemes also can be improved using our framework.

### 3.4 Application of Image Authentication

There are many applications for perceptual image hashing. For example, the hash distance between an original image and its tampered version should be great enough for the application of image authentication. This is helpful to determine whether a suspicious image has been tampered or not. Figure 7 shows several cases of image tampering. The top of each image pair is the original image while the bottom is the tampered version. Table 4 lists the hash distances between the image pairs.



**Fig. 7** Cases of image tampering

**Table 4** Hash distances between original and tampered images

Schemes	Figure 6a	Figure 6b	Figure 6c
RE	0.025714	0.00036743	0.029533
RE-P	<b>0.025811</b>	<b>0.00040721</b>	<b>0.030010</b>
RP-NMF	0.0026688	0.00017152	0.014278
RP-NMF-P	<b>0.0030037</b>	<b>0.00021065</b>	<b>0.015839</b>
DAE-NN	0.0015845	0.00038946	0.0034646
DAE-NN-P	<b>0.0016139</b>	<b>0.00039232</b>	<b>0.0034675</b>

The bold stand for the better performances

It is observed from the results that hash distances of our framework are greater than that of existing hash schemes. The results indicate that our framework is more effective to capture the changes caused by tampering manipulations. Therefore, our framework can achieve better performance for the application of image authentication.

## 4 Conclusion

With more and more perceptual image hashing methods are designed, we investigate the improvement of perceptual image hashing instead of developing a new hashing method. We propose a general framework to improve existing image hashing schemes using the strategy of reference image construction. Existing image hashing schemes are executed on the input image and reference image, respectively. The obtained

hashes of input image and reference image are combined to form a new hash sequence. Experimental results show that the performances of the new hash sequence, e.g., perceptual robustness, discrimination capability, are better than that of the original ones. In future work, more ways (such as concatenation) of fusing the last two hash sequences can be discussed for improving perceptual image hashing further.

**Acknowledgements** This work was supported by Natural Science Foundation of China under Grant 62002214.

## References

- Chuan, Q., Chinchen, C., Peiling, T.: Robust image hashing using non-uniform sampling in discrete Fourier domain. *Digital Signal Process.* **23**(2), 578–585 (2013)
- Xun, Y.: Hash function based on chaotic tent maps. *IEEE Trans. Circ. Syst. **52***(6), 354–357 (2005)
- Yan, Z., Shuzhong, W., Xinpeng, Z., Heng, Y.: Robust hashing for image authentication using zernike moments and local features. *IEEE Trans. Inf. Forensics Secur.* **8**(1), 55–63 (2013)
- Hongjia, Z., Shenqi, L., Hanyang, J., Xueming, Q., Tao, M.: Deep transfer hashing for image retrieval. *IEEE Trans. Circ. Syst. Video Technol.* **31**(2), 742–753 (2021)
- Caiping, Y., Chiman, P., Xiaochen, Y.: Quaternion-based image hashing for adaptive tampering localization. *IEEE Trans. Inf. Forensics Secur.* **11**(12), 2664–2677 (2016)
- Lefèvre, F., Macq, B., Legat, J.: RASH: RAdon soft hash algorithm. In: 11th European Signal Processing, pp. 1–4. IEEE, Piscataway, NJ, USA (2002)
- Ling, D., Anthony, T., Rummim, C.: Perceptual hashing for image authentication: a survey. *Sig. Process. Image Commun.* **81**(115713) (2020)
- Mehran, K., Kave, E., Bir, B.: Discrete cosine transform locality-sensitive hashes for face retrieval. *IEEE Trans. Multimedia* **16**(4), 1090–1103 (2014)
- Xin, Z., Fuchun, S., Guangcan, L., Yi, M.: Fast low-rank subspace segmentation. *IEEE Trans. Knowl. Data Eng.* **26**(5), 1293–1297 (2014)
- Liwei, K., Chunshien, L., Chaoyung, H.: Compressive sensing-based image hashing. In: 16th International Conference on Image Processing (ICIP), pp. 1285–1288. IEEE, New York, USA (2009)
- Zhenjun, T., Xianquan, Z., Liyan, H., Yumin, D.: Robust image hashing using ring-based entropies. *Signal Process.* **93**(7), 2061–2069 (2013)
- Zhenjun, T., Xianquan, Z., Shichao, Z.: Robust perceptual image hashing based on ring partition and NMF. *IEEE Trans. Knowl. Data Eng.* **26**(3), 711–724 (2014)
- Zhenjun, T., Xianquan, Z., Xianxian, L., Shichao, Z.: Robust image hashing with ring partition and invariant vector distance. *IEEE Trans. Inf. Forensics Secur.* **11**(1), 200–214 (2016)
- Tianqiang, P., Fang, L.: Image retrieval based on deep convolutional neural networks and binary hashing learning. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1742–1746. IEEE, New York, USA (2017)
- Yuenan, L., Dongdong, W., Linlin, T.: Robust and secure image fingerprinting learned by neural network. *IEEE Trans. Circuits Syst. Video Technol.* **30**(2), 362–375 (2020)
- Chuan, Q., Enli, L., Guorui, F., Xinpeng, Z.: Perceptual image hashing for content authentication based on convolutional neural network with multiple constraints. *IEEE Trans. Circuits Syst. Video Technol.* (2020). <https://doi.org/10.1109/TCSVT.2020.3047142>
- Zichi, W., Zhenxing, Q., Xinpeng, Z., Sheng, L.: An improved steganalysis method using feature combinations. In: Xinming, S., Zhaoqing, P., Bertino, E. (eds.) 5th International Conference on Artificial Intelligence and Security, pp. 115–127. Springer international publishing, Cham, Switzerland (2019)

18. Kutter, M., Winkler, S.: A vision-based masking model for spread-spectrum image watermarking. *IEEE Trans. Image Process.* **11**(1), 16–25 (2002)
19. Zichi, W., Zhenxing, Q., Xinpeng, Z., Min, Y., Dengpan, Y.: On improving distortion functions for JPEG steganography. *IEEE Access* **6**(1), 74917–74930 (2018)
20. Schaefer, G., Stich, M.: UCID—An uncompressed colour image database. In: Yeung, M.M., Lienhart, R.W., Li, C.S. (eds.) Conference on Storage and Retrieval Methods and Applications for Multimedia, pp. 472–480. SPIE, Bellingham, USA (2004)

# A Zero-Watermarking Against Large-Scale Cropping Attack



Jing Wang , Sellappan Palaniappan , and Bing He

**Abstract** Considering the digital copyright protection of the color image, and aiming at the shortcomings of the existing color image zero-watermarking algorithms against various attacks including image information loss, e.g., cropping, rows and columns removal, and smearing, etc. A color image zero-watermarking algorithm based on quaternion fractional-order generalized Laguerre moments (QFr-GLMs) is proposed. The novelties of the proposed scheme are as follows: (1) QFr-GLMs are applied the first time in the digital copyright protection based on color image zero-watermarking scheme; (2) The proposed zero-watermarking scheme can efficiently resist traditional noisy and smoothing filter attacks, moreover, when the image is attacked by large-scale cropping or smearing, the proposed zero-watermarking scheme can still detect the watermark correctly; Simulation experimental results prove that the performance of the proposed scheme is superior to that of the graying-based and the single channel-based zero-watermarking schemes.

**Keywords** Quaternion algebra · Fractional-order generalized Laguerre moments · Zero-watermarking · Cropping attack

## 1 Introduction

As an effective image descriptor, orthogonal moments have been widely used in the fields of image processing and analysis such as image reconstruction, digital image watermarking and objection recognition. Orthogonal moments can be divided into continuous orthogonal moments like Zernike moments [1], orthogonal Fourier-Mellion moments [2] and Bessel-Fourier moments [3], or into discrete orthogonal moments. Compared with continuous orthogonal moments, the discrete orthogonal moments have better numerical stability because the image itself is digitalized and

---

J. Wang · S. Palaniappan

Malaysia University of Science and Technology, 47810 Petaling Jaya, Malaysia  
e-mail: [wang.jing@phd.must.edu.my](mailto:wang.jing@phd.must.edu.my)

J. Wang · B. He

Weinan Normal University, Weinan 714000, China

there is no numerical approximation operation. In the regions of image processing and analysis, color images include more information than grayscale and binary images. Quaternion algebra based color image representation regards an image as a three-dimensional vector describing the components of the color image, which effectively uses the color information of different channels of the color image. For color images, quaternion algebra can be used to construct quaternion image moments. Many quaternion image moments, such as: quaternion polar harmonic Fourier moments (QPHFMs) [4], and quaternion exponent-Fourier moments (QEPMs) [5], have been proposed by researchers using the existing orthogonal polynomials. In order to make the image description more stable and accurate, the related scholars extend the integer-order image moments to the fractional-order level, and propose the related fractional-order orthogonal moments, such as the fractional-order Chebyshev moments (Fr-CMs) [6], and fractional-order Zernike moments (Fr-ZMs) [7].

Nowadays, digital copyright protection of image content is a key issue to protect the rights of intellectual property for authors, due to those digital image data can be easily carried out manipulations. Digital watermarking is considered a vital copy-right protection technique by embedding watermark (or a message) into the images. At present, in the field of digital watermarking, many scholars and researchers have focused on the research of zero watermarking. In recent years, zero watermarking technology has become a research hot-spot in the field of information security, due to it solves the contradiction between invisibility and robustness in traditional watermarking technology. Wen et al. [8] have proposed the concept of zero watermark for the first time in 2003, since then, the research on zero watermarking has been continuously expanded and enhanced, and more and more related academic research results have been applied in the field of digital copyright.

Inspired by the Fractional-order Generalized Laguerre Moments (Fr-GLMs) that can effectively extract local features of an image, in our work, we propose a novel zero-watermarking scheme, to protect the copyright of color images, based on Quaternion Fractional-order Generalized Laguerre Moments (QFr-GLMs). This scheme takes advantage of Region of Interest (ROI) feature-extraction of the QFr-GLMs to improve the robustness against large-scale cropping and smearing attacks in the process of image transmission. In addition, since the technique of quaternion algebra is introduced, compared with the graying-based and the single channel-based zero-watermarking schemes, the performance of our proposed scheme is more excellent in resisting conventional image processing attacks.

## 2 Definition and Calculation of QFr-GLMs

Firstly, the three components of a color image  $f^{rgb}(x, y)$ ,  $f_r(x, y)$ ,  $f_g(x, y)$ , and  $f_b(x, y)$ , correspond to three imaginary components of a pure quaternion. Therefore, an image  $f^{rgb}(x, y)$  in RGB color space can be expressed by the following quaternion.

$$f^{rgb}(x, y) = f_r(x, y)i + f_g(x, y)j + f_b(x, y)k. \quad (1)$$

Secondly, the corresponding fractional-order generalized Laguerre moments (Fr-GLMs) can be defined as

$$S_{nm}^{(\alpha, \lambda)} = w \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f^{gray}(i, j) \bar{L}_n^{(\alpha_x, \lambda_x)}(x_i) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_j), \quad (2)$$

where  $f^{gray}(i, j)$  represents a grayscale digital image. For convenience, we map the original two-dimensional digital-image matrix to a square area of  $[0, L] \times [0, L]$ . Here,  $L > 0$ ,  $w = (L/N)^2$ ,  $x_i = \frac{iL}{N}$ ,  $y_j = \frac{jL}{N}$ ,  $i, j = 0, 1, 2, \dots, N-1$ , and the normalized fractional-order generalized Laguerre polynomial (NFr-GLPs)  $\bar{L}_n^{(\alpha, \lambda)}(x)$  can be recursively calculated as follows

$$\bar{L}_n^{(\alpha, \lambda)}(x) = (A_0 + A_1 x^\lambda) \bar{L}_{n-1}^{(\alpha, \lambda)}(x) + A_2 \bar{L}_{n-2}^{(\alpha, \lambda)}(x), \quad (3)$$

where  $\bar{L}_0^{(\alpha, \lambda)}(x) = \sqrt{\frac{\omega^{(\alpha, \lambda)}(x)}{\Gamma(\alpha+1)}}$ ,  $\bar{L}_1^{(\alpha, \lambda)}(x) = (1 + \alpha - x^\lambda) \sqrt{\frac{\omega^{(\alpha, \lambda)}(x)}{\Gamma(\alpha+2)}}$ ,  $A_0 = \frac{2n+\alpha-1}{\sqrt{n(n+\alpha)}}$ ,  $A_1 = \frac{-1}{\sqrt{n(n+\alpha)}}$ ,  $A_2 = -\sqrt{\frac{(n+\alpha-1)(n-1)}{n(n+\alpha)}}$ ,  $\omega^{(\alpha, \lambda)}(x) = \lambda x^{(\alpha+1)\lambda-1} \exp(-x^\lambda)$ ,  $\alpha > -1$ ,  $\lambda > 0$ ,  $\lambda \in R^+$ , and  $\Gamma(\cdot)$  is the gamma function, for details, see Ref. [9].

Finally, the right-side QFr-GLMs of an original RGB color image in Cartesian coordinates are defined as

$$\begin{aligned} Q_{nm}^{(\alpha, \lambda)} &= w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) f^{rgb}(p, q) \mu \\ &= -\frac{1}{\sqrt{3}} \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (f_r + f_g + f_b) \right] \\ &\quad + \frac{1}{\sqrt{3}} k \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (f_r - f_g) \right], \\ &\quad + \frac{1}{\sqrt{3}} j \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (f_b - f_r) \right] \\ &\quad + \frac{1}{\sqrt{3}} i \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (f_g - f_b) \right] \end{aligned} \quad (4)$$

where  $\mu = (i + j + k)/\sqrt{3}$  is the unit pure imaginary quaternion. The QFr-GLMs expressed in quaternions and the Fr-GLPs of single channels in traditional RGB color images are related as follows:

$$Q_{nm}^{(\alpha,\lambda)} = A + iB + jC + kD, \quad (5)$$

$$\begin{aligned} A &= -\frac{1}{\sqrt{3}}[S_{nm}^{(\alpha,\lambda)}(f_r) + S_{nm}^{(\alpha,\lambda)}(f_g) + S_{nm}^{(\alpha,\lambda)}(f_b)], \\ B &= \frac{1}{\sqrt{3}}[S_{nm}^{(\alpha,\lambda)}(f_g) - S_{nm}^{(\alpha,\lambda)}(f_b)], C = \frac{1}{\sqrt{3}}[S_{nm}^{(\alpha,\lambda)}(f_b) - S_{nm}^{(\alpha,\lambda)}(f_r)], \\ D &= \frac{1}{\sqrt{3}}[S_{nm}^{(\alpha,\lambda)}(f_r) - S_{nm}^{(\alpha,\lambda)}(f_g)]. \end{aligned}$$

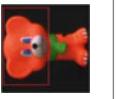
Accordingly, an original color image  $f^{rgb}(p, q)$  can be reconstructed by finite-order QFr-GLMs. The reconstructed image is represented as

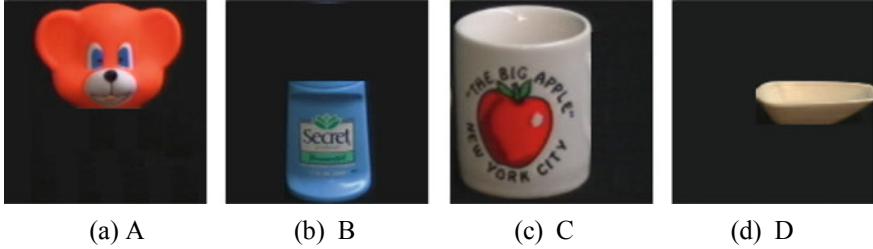
$$\begin{aligned} \bar{f}^{rgb}(p, q) &= w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} Q_{nm}^{(\alpha,\lambda)} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) \mu \\ &= -\frac{1}{\sqrt{3}} \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (B + C + D) \right] \\ &\quad + \frac{1}{\sqrt{3}} k \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (A + B - C) \right] \\ &\quad + \frac{1}{\sqrt{3}} j \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (A - B + D) \right] \\ &\quad + \frac{1}{\sqrt{3}} i \left[ w \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \bar{L}_n^{(\alpha_x, \lambda_x)}(x_p) \bar{L}_m^{(\alpha_y, \lambda_y)}(y_q) (A + C - D) \right] \quad (6) \end{aligned}$$

### 3 Local-Feature-Extraction Analysis for QFr-GLMs

In this section, we develop the application of the QFr-GLMs to local-feature extraction from color images and the local feature image-cropping invariance of the QFr-GLMs is tested. The test images were four typical color images selected from the COIL-100 database, i.e., Toy, Bottle, Cup and Spoon size of  $128 \times 128$ , respectively. The local features in each color image at different positions of the four-color images were reconstructed using the features extracted by the QFr-GLMs with different parameters. The experimental results are summarized in Table 1. This table shows that under different parameter settings, the proposed QFr-GLMs provided good image reconstructions in different regions of the original color image (the target areas of ROI extraction from the original color images are enclosed in the red-edged boxes). Under the parameter setting  $\alpha_x = 22, \alpha_y = 1, \lambda_x = 1.4, \lambda_y = 1.6$ , the QFr-GLMs extracted the upper part of the original color image. Meanwhile, the QFr-GLMs with  $\alpha_x = 3, \alpha_y = 60, \lambda_x = 1.21, \lambda_y = 1.25$  extracted the bottom part of the original color image, those with  $\alpha_x = 5, \alpha_y = 3, \lambda_x = 1.46, \lambda_y = 1$  obtained the left part

**Table 1** Local-image reconstruction performances of the QFr-GLMs

Original images				
The selected ROI images from original images				
The ROI reconstructed images				
Parameter values	$\alpha_x = 22, \alpha_y = 1,$ $\lambda_x = 1.40, \lambda_y = 1.60$ $L = 30, n = m = 20$	$\alpha_x = 3, \alpha_y = 60,$ $\lambda_x = 1.21, \lambda_y = 1.25$ $L = 30, n = m = 20$	$\alpha_x = 5, \alpha_y = 3,$ $\lambda_x = 1.46, \lambda_y = 1$ $L = 30, n = m = 18$	$\alpha_x = 1, \alpha_y = 100,$ $\lambda_x = 1.28, \lambda_y = 1.38$ $L = 30, n = m = 18$



**Fig. 1** The color cropped images at different positions from coil-100 datasets

**Table 2** QFr-GLMs moduli and MRE for color cropped images at different positions

MRE	$ Q_{11}^{(\alpha,\lambda)} $	$ Q_{22}^{(\alpha,\lambda)} $	$ Q_{33}^{(\alpha,\lambda)} $	$ Q_{44}^{(\alpha,\lambda)} $	$ Q_{55}^{(\alpha,\lambda)} $	$ Q_{66}^{(\alpha,\lambda)} $	$ Q_{77}^{(\alpha,\lambda)} $	MAE
Toy	0.4135	1.9333	0.5847	1.4608	0.6005	0.8636	0.7116	–
A	0.4135	1.9333	0.5847	1.4608	0.6005	0.8636	0.7116	0
Bottle	0.3435	0.2851	0.6614	1.5392	0.7979	0.1921	0.3510	–
B	0.3435	0.2851	0.6613	1.5389	0.7972	0.1909	0.3485	0.0007
Cup	0.2371	1.4215	0.7482	1.6729	1.0414	0.6900	0.2739	–
C	0.2371	1.4215	0.7482	1.6729	1.0414	0.6900	0.2739	0
Spoon	0.0511	3.1356	0.0293	1.1299	0.1897	0.0869	0.0245	–
D	0.0511	3.1356	0.0291	1.1297	0.1903	0.0887	0.0246	0.0004

of the original color image, and those with  $\alpha_x = 85, \alpha_y = 3, \lambda_x = 1.46, \lambda_y = 1.5$  extracted the right-upper part of the original color image. In addition, the image-cropping invariance of the QFr-GLMs is tested, and each of four-color images is cropped at different positions, respectively, which is illustrated in Fig. 1. The moduli for the QFr-GLMs of the four-color images are compared with those for the QFr-GLMs of the cropped images. The change rate of the moduli of the rotated images to that of the original images is measured by mean absolute error (MAE). The results in Table 2 shows that all MAE values are smaller than 0.0007, indicating that the moduli for the QFr-GLMs of the cropped images remain basically unchanged, which demonstrates that the QFr-GLMs under different parameter settings are invariant to image-cropping.

## 4 Proposed Zero-Watermarking Scheme

In this section, the QFr-GLMs were applied to zero-watermarking color image. Although the existing anti-geometric attack watermarking algorithms are effective against traditional geometric transformations (rotations, scaling, translations, or affine transformations) and conventional signal processing (noise adding, lossy

compression, and image filtering), the information loss from strong cropping and shearing attacks have been rarely investigated in this field. Some of the robust watermarking algorithms can extract the watermark information to a certain extent after weak cropping (information loss < 25%), but cannot recover the watermark information after large-scale information loss (strong cropping or shear attacks). In such cases, the algorithms are invalid. To resolve these problems, we exploit the advantages of the proposed QFr-GLMs (local-feature extraction from color images and multi-point embedding) in a zero-watermarking system corrupted by large-scale cropping and smearing attacks.

#### **4.1 Design of the Zero-Watermarking Scheme**

The registration process of the proposed color-image zero-watermarking algorithm is shown as follows:

Step 1: Divide the original color host image needing copyright protection into  $W \times W$  blocks. To preserve the details of the block image and reduce the false-alarm rate in digital watermarking, the size of the block image segmented in the host image should not be too small. Empirically, the size of the block image (if possible) should exceed 1/9 of the original image size. In this paper, the color host image  $I_{M \times N}^{(RGB)}$  was sized  $192 \times 192$ . Therefore, we divided the original image into  $3 \times 3$  blocks, and marked the segmented images as  $B_k^{(RGB)}$ .

Step 2: Extract the local-feature parameters (i.e.,  $\alpha_x$ ,  $\alpha_y$ ,  $\lambda_x$ ,  $\lambda_y$ , and  $L$ ) of the color host image corresponding to each block image, and set them as the key character sequence  $key_i = \{\alpha_x^i, \alpha_y^i, \lambda_x^i, \lambda_y^i, L^i\}$ ,  $i = 1, 2, \dots, 9$ . A number of lower-order moments of the proposed QFr-GLMs are then extracted as the feature vector  $V_k^{(RGB)}$ ,  $k = 1, 2, \dots, 9$ . Here we set  $V_k^{(RGB)} = \{v_{nm}^{(1)}, v_{nm}^{(2)}, \dots, v_{nm}^{(k)}, k = 9\}$ ,  $v_{nm}$  denote the lower-order moments of  $Q_{nm}^{(\alpha, \lambda)}$  in Eq. (4). This feature vector is registered in the Certificate Authority Center (CA center, a third-party copyright protection center).

Step 3: The above information is time stamped and registered in the CA center, together with the user's signature information. At this time, the original color host image is announced to be under copyright protection.

#### **4.2 Design of the Zero-Watermarking Detection System**

Watermark detection is the reverse process of registration, in this stage, extracting the feature vector  $\tilde{V}_k^{(RGB)}$  from the CA center, and sum the absolute differences between the feature vector  $\tilde{V}_k^{(RGB)}$  obtained in the previous step and the feature vector  $V_k^{(RGB)}$  obtained in the registration stage of the zero-watermarking algorithm. The sum is calculated as follows:

$$d = \min \left\{ \frac{1}{cnt^2} \sum_{n=1}^{cnt} \sum_{m=1}^{cnt} |v_{nm}^{(k)} - \tilde{v}_{nm}^{(k)}| \right\}. \quad (7)$$

In this experiment,  $cnt = 10$  and  $k = 1, 2, \dots, 9$ . When  $d \geq \varepsilon$  (where  $\varepsilon$  is an empirical threshold, set to 0.02 in the current experiment) and the time stamp does not match the information provided by the CA center, the verification is completed. The verification proves that the color host image either contains or lacks the watermark information.

### 4.3 Experimental Results of the Proposed Scheme

To verify the effectiveness of the zero-watermarking algorithm based on the proposed QFr-GLMs, the “cat” color image was extracted from the COIL-100 database, and its size was normalized to  $192 \times 192$ . The size-normalized image was taken as the host image (see Fig. 2). This subsection describes the implementation and results of two groups of experiments. The first experiment investigated the robustness of the algorithm to various signal-processing. The second experiment examined the performance of the proposed zero-watermarking system on strongly cropped and randomly smeared images, and visualized the results of different cropping and smearing ratios on the cat host color image. In this subsection, the similarity between the original color host image and the attacked color image was measured by the  $PSNR$ , and the similarity between the detected feature vector  $\tilde{V}_k^{(RGB)}$  and the feature vector  $V_k^{(RGB)}$  in the CA center was judged by Eq. (4). Evidently, as  $d$  approaches 0, the zero-watermarking scheme becomes more robust.

**Fig. 2** Original color host image of a cat ornament



**Experiment 1.** During network transmission, images are highly vulnerable to noise interference. In this experiment, the original color host image was tested under attacks by Gaussian white noise with a variance of 0.02, salt and pepper noise with a density of 5%, and speckle noise with a variance of 0.04. As noise interference is usually removed by filtering operations, filtering (with consequent loss of image pixels) is another frequently encountered attack in image processing and pattern recognition. In the experiment, the original color host image was subjected to median and average filtering with a  $5 \times 5$  window. Furthermore, images in computer vision or signal processing are frequently processed by blurring and JPEG compression. Image blurring leads to partial block distortions, and JPEG compression causes pixel losses. During this experiment, the Gaussian blur parameter was set to the default in Matlab2013a's software environment, and the JPEG compression ratio was set to 60. Table 3 compares the results of the proposed zero-watermarking scheme and other schemes (the direct graying-based method and the single channel-based method). The proposed algorithm was highly robust and outperformed the two established schemes.

**Experiment 2.** Cropping or smearing attacks are the most difficult problems in digital watermarking. Although the existing digital watermarking algorithms are robust to small-scale cropping and smearing operations, most of them fail under large-scale cropping or smearing operations. In this experiment, the color host image was cropped by different ratios (25%, 70%, 68%, and 82%) on different regions of the image. The image was also randomly smeared at ratios of 16%, 74%, and

**Table 3** Typical results of images subjected to traditional image processing

Attacks	PSNR (dB)	$d$		
		Direct graying-based method [11]	Single channel-based method [12]	The proposed method
Gaussian white noise (Variance: 0.02)	17.82	0.0024	0.00157	0.00112
Salt and peppers noise (Density: 15%)	15.63	0.0043	0.00172	0.00121
Speckle noise (Variance: 0.04)	16.92	0.0025	0.00168	0.00099
Median filtering (Window: 5)	26.43	0.00038	0.00021	0.00018
Average filtering (Window: 5)	27.68	0.00022	0.00016	0.00011
Gaussian blur (Default)	28.22	0.00018	0.00013	0.00008
JPEG compression (Quality: 60%)	25.44	0.00029	0.00019	0.00014

86%. Table 4 shows the cropped images and their experimental results. Unlike the zero-watermarking algorithm of Xia et al. [10], the proposed scheme based on the QFr-GLMs effectively detected the watermark, even on the image cropped by 82%. Thus, the proposed scheme is suitable for an image-copyright protection of cropped or smeared images.

## 5 Conclusions

This paper proposed a new zero-watermarking scheme, to protect color image copyright, based on QFr-GLMs. Our proposed scheme is very robust against traditional image processing, such as: noise, and smoothing filtering attacks, especially, against large-scale cropping and smearing attacks in the fields of image watermarking, which provide significant practical value in the copyright protection of color images. In future work, we would aim to address other research fields related to color images such as image retrieval, image classification, image matching and other research fields.

**Funding** This work was supported by the Science Foundation of the Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing (Grant No. XUPT-KLND201901), Shaanxi Province of Key R&D project (Grant No. 2020GY-051), Weinan regional collaborative innovation development research project (Grant No. WXQY001-001 and WXQY002-007) and Shaanxi Provincial Department of Education project (Grant No. 20JS044).

**Table 4** Experimental results of images subjected to cropping and smearing attacks

Host image				
Attacks	Parameters	Attacked color image	$d$	Detection results (Y/N)
Cropping	Upper-left cropping area: 25%		0	Y
Cropping	Rows cropping area: 70%		0.0082	Y
Cropping	Columns cropping area: 68%		0.0076	Y
Cropping	Upper-left Cropping area: 82%		0.0097	Y

(continued)

**Table 4** (continued)

Host image	 $d = 0$			
Attacks	Parameters	Attacked color image	$d$	Detection results (Y/N)
Smearing	Smearing area: 16%		0.0007	Y
Smearing	Smearing area: 74%		0.0128	Y
Smearing	Smearing area: 86%		0.03862	N

## References

1. Singh, C., Ranade, S.K.: Image adaptive and high-capacity watermarking system using accurate Zernike moments. *IET Image Proc.* **8**(7), 373–382 (2014)
2. Singh, C., Upneja, R.: Accurate computation of orthogonal Fourier-Mellin moments. *J. Math. Imaging Vis.* **44**(3), 411–431 (2012)
3. Yang, T., Ma, J., Miao, Y., et al.: Quaternion weighted spherical Bessel-Fourier moment and its invariant for color image reconstruction and object recognition. *Inf. Sci.* **505**, 388–405 (2019)
4. Wang, C., Wang, X., Li, Y., et al.: Quaternion polar harmonic Fourier moments for color images. *Inf. Sci.* **450**, 141–156 (2018)
5. Wang, C.P., Wang, X.Y., Xia, Z.Q., et al.: Geometrically resilient color image zero-watermarking algorithm based on quaternion exponent moments. *J. Vis. Commun. Image Represent.* **41**, 247–259 (2016)
6. Benouini, R., Batioua, I., Zenkouar, K., et al.: Fractional-order orthogonal Chebyshev moments and moment invariants for image representation and pattern recognition. *Pattern Recogn.* **86**, 332–343 (2019)
7. Chen, B., Yu, M., Su, Q., et al.: Fractional quaternion Zernike moments for robust color image copy-move forgery detection. *IEEE Access* **6**, 56637–56646 (2018)
8. Wen, Q., Sun, T.F., Wang, S.X.: Concept and application of zero-watermark. *Acta Electronica Sin.* **31**(2), 214–216(2003)
9. Bhrawy, A.H., Alhamed, Y.A., Baleanu, D., et al.: New spectral techniques for systems of fractional differential equations using fractional-order generalized Laguerre orthogonal functions. *Fractional Calc. Appl. Anal.* **17**(4), 1137–1157 (2014)
10. Xia, Z.Q., Wang, X.Y., Zhou, W.J., et al.: Color medical image lossless watermarking using chaotic system and accurate quaternion polar harmonic transforms. *Signal Process.* **157**, 108–118 (2019)
11. Jin, N.R., Lv, X.Q., Gu, Y., et al.: Blind watermarking algorithm for color image in Contourlet domain based on QR code and chaotic encryption. *Packag. Eng.* **38**(15), 173–178 (2017)
12. Hong, Y.L., Qie, G.L., Yu, H.W., et al.: A contrast preserving color image graying algorithm based on two-step parametric subspace model. *Front. Inf. Technol. Electron. Eng.* **11**, 102–116 (2017)

# A Brief Review of Image Dehazing Algorithms Based on Deep Learning



Juan Wang , Chang Ding , Minghu Wu , Yuanyuan Liu ,  
and Guanhui Chen

**Abstract** Single image dehazing is a challenging problem, and it is far from solved. The application of deep learning in dehazing is only in the initial stage of exploration since the structure of deep learning is not designed for it. It occurs frequently that outdoor image quality is seriously affected when capturing image outside with dense haze, the contrast of the picture drops, and the information is lost due to the particles in the atmosphere. It seems indispensable to work on images without dehazing. A great number of methods have been proposed over the past dozen years. Those methods can be divided into traditional and deep learning methods. **This paper mainly summarizes and uses traditional algorithms to compare, explains the classic algorithms of deep learning and introduces recent new efficient algorithms.** The deep learning method architecture in the paper has been classified into the following two categories, (a) Convolution neural network (CNN) and (b) Generative adversarial network (GAN).

**Keywords** Dehazing · Deep learning · Convolution neural network · Generative adversarial network

## 1 Introduction

When light propagates in haze and other media, the reflected light of the target object encounters these particles during the propagation process and interacts with them, such as absorption, radiation, and scattering, resulting in the redistribution of light energy, or reduced contrast in the color distortion and other phenomena.

Since insufficient information, these low-quality images seriously affect the effectiveness of the intelligent vision system. Hazy images will affect the target observation, identification and forensics. Due to the scattering of particles, event analysis

---

J. Wang · C. Ding · M. Wu ( ) · Y. Liu · G. Chen

Hubei Energy Internet Engineering Technology Research Center, Hubei University of Technology, Wuhan 430068, China

Hubei Laboratory of Solar Energy Efficient Utilization and Energy Storage Operation Control, Hubei University of Technology, Wuhan 430068, China

becomes difficult, and the image information collected by the imaging sensor is severely degraded, which greatly limits the application value of the image.

The purpose of image dehazing is to eliminate or alleviate the impact of haze environment on image quality, and enhance the readability of images. Based on dehazing technique, knowledge in every direction in intelligent vision can be applied to smart vehicles, parking and other intelligent image recognition situations. What makes the field of image dehazing a hotspot is the proposal of these methods, such as Dark Channel Prior (DCP) [1], Maximum Contrast [2] (MC). Dehazing algorithms based on prior knowledge are less proposed recently [3].

Traditional dehazing model is based on prior knowledge, but it does not perform well in heavily hazy images or complicated situations. The use conditions of traditional dehazing models are relatively strict, and programs require long running time.

Deep learning methods have solved these key problems. DehazeNet [4], early effective deep learning methods, adopts convolutional neural network for single image haze removal and system used to restore clear images.

CNN-based methods, like fusion work [5, 6], Multi-domain application [7, 8], methods based on GAN architecture are [9–12]. And dehazing algorithms are used in many scenarios, such as underwater circumstances [13–15], nighttime dehazing [16–18], nonhomogeneous haze image [19–21].

## 2 Related Work

$$\mathbf{I}(x) = \mathbf{J}(x)\mathbf{t}(x) + \mathbf{A}(1 - \mathbf{t}(x)), \mathbf{t}(x) = e^{-\beta d(x)} \quad (1)$$

Atmospheric Scattering Model [22] was widely used to describe the formation of hazy image. Where  $\mathbf{I}(x)$  represents haze image,  $\mathbf{J}$  denotes the scene radiance which represents the haze-free image,  $\mathbf{A}$  denotes the global atmospheric light, and  $\mathbf{t}$  denotes the medium transmission.  $\mathbf{t}$  is defined as follows if the global atmosphere is homogeneous.

$\beta$  stands for scattering coefficient of atmosphere, and  $d$  is scene depth, the atmosphere scattering model shows that image dehazing is a determined problem with the knowledge of  $\mathbf{A}$  and  $\mathbf{t}$ .

This paper lists several image datasets commonly used in the dehazing community (Table 1).

## 3 Traditional Algorithms

**DCP.** He et al. purposed Dark Channel Prior method DCP. They found that the non-sky local area image in the test image always has at least one low-value channel, which means the minimum value of light intensity in this area has a minimum value

**Table 1** Available datasets and information

Datasets	In/outdoor	Format	Haze
NYU Depth Dataset V2 [23]	Indoor	JPG	Synthetic
FRIDA [24]	Outdoor	PNG	Synthetic
D-HAZY [25]	Indoor	PNG	Synthetic
O-HAZE [26]	Outdoor	PNG	Artificial
I-HAZE [27]	Indoor	PNG	Artificial
RESIDE [28]	Both	PNG	Synthetic
BeDDE [29]	outdoor	PNG	Natural

(Fig. 1).

$$J^{dark}(x) = \min_{c \in \{r, g, b\}} (\min_{y \in \Omega(x)} (j^c(y))) \quad (2)$$

$J_C$  represent the color channel of  $\mathbf{J}$ ,  $\Omega(\mathbf{x})$  denotes a local patch center at  $\mathbf{x}$ . Intensity of  $J_{dark}$  tends to be zero, called dark channel.

The minimum operation is performed on the three color channels separately, DCP supposed  $t$  is constant, and parameter  $A$  is a fixed value. Minimize both sides of the equation. According to the theory dark channel prior,  $J_{dark}$ , haze-free image is approximately equal to zero. There are few particles in the atmosphere, it feels unreal if it looks completely clear. Keeping some haze image vision will increase authenticity. Parameter  $\omega$  value is between 0 and 1, and DCP takes 0.95 as their fundamental value.

DCP set the typical value of  $t_0$  as 0.1, and the dehazing image looked dark since the scene radiance is usually not as bright as the atmospheric light, and recovery formula:

$$J(x) = \frac{I(x) - A}{\max(t(x), t_0)} + A \quad (3)$$



(a) Haze image

(b) Dark channel

**Fig. 1** Hazy image and corresponding dark channel map

After more experiments, it was found that the dehaze image does not perform well in certain circumstances, such as mismatch between different parts of an image, or some areas of the image are not restored entirely, since the transmission  $t$  is not accurate enough.

## 4 Deep Learning Algorithms

Recently, researchers are beginning to try to replace traditional image dehazing methods with deep learning-based methods. There are two main methods to dehazing, one is based on the atmospheric degradation model and the neural network is used to estimate the parameters in the model. Most of the early methods are based on this idea. The other one is to use the input haze image, directly output the image after processed, the latest dehazing methods currently favor the latter.

### 4.1 Convolution Neural Network

**AOD-Net.** AOD-Net [30] proposed an end-to-end model based on CNN, the network could generate the clear image directly instead of calculating parameters and bring the parameters back into the formula. Which means all the parameters could be calculated in unit model. Equation (1) is the foundation of the atmospheric scattering model, parameters  $t(x)$  and  $A$  are centralized into one equation, and the equation transform into:

$$J(x) = K(x)I(x) - K(x) + b, K(x) = \frac{\frac{1}{t(x)}(t(x) - A) + (A - b)}{t(x) - 1} \quad (4)$$

The value of  $b$  is constant at 1, therefore, our goal is to set an input adaptive depth model whose parameters will change with the input haze image, thereby minimizing the reconstruction error between the output  $J(x)$  and the ground truth clear image since  $K(x)$  depends on  $I(x)$ .

AOD-NET is not a heavy structural levels network, and the processing time of an image only takes one third time of DehazeNet, benefiting from its convolutional layer containing only three kernels. AOD used ReLU rather than BReLU, and the learning rate is 0.001. Comparing the result of AOD-NET and previous research.

The lightweight of AOD-Net and its characteristics as the integrated dehazing model made the AOD method processing speed much faster.

**DCPDN.** An end-to-end dehazing algorithms which are enabled by Embedded method use transmission map and atmospheric light to calculate the residual haze block directly through network. DCPDN [31] adopted edge-preserving encoder with

densely connected pyramids, and it was used to accurately estimate transmission mapping.

The novelty is that the DCPDN uses the joint discriminator in the GAN framework to determine whether the paired samples come from the data distribution.

Network consists of four modules: (a) Pyramid densely connected transmission graph estimation network. (b) Atmospheric light estimation network. (c) Formula embedding. (d) Joint discriminator network. DCPDN used dense blocks to extract image features and update to the original size after pooling four pictures of different sizes.

DCPDN supposed the atmospheric light is uniform. Estimating the atmospheric light of the image through U-net, bringing it and transmission map into equation.

$$\hat{J}(z) = \frac{I(z) - \hat{A}(z)(1 - \hat{t}(z))}{\hat{t}(z)} \quad (5)$$

DCPDN defined a new loss function:

$$L^E = \lambda_{E,l2} L_{E,l2} + \lambda_{E,g} L_{E,g} + \lambda_{E,f} L_{E,f} \quad (6)$$

Which consists of edge-preserving loss  $L_E$ , two-directional gradient loss,  $L_2$  loss  $L_E$ ,  $L_{E,l2}$ . and feature loss  $L_{E,f}$ .

DCPDN proposed the joint-discriminator which is defined as follow:

$$\begin{aligned} & \min_{G_t, G_d} \max_{D_{joint}} E_{I \sim P_{data(I)}} [\log(1 - D_{joint}(G_t(I)))] + E_{I \sim P_{data(I)}} [\log(1 - D_{joint}(G_d(I)))] \\ & + E_{t, I \sim P_{data(t, J)}} [\log D_{joint}(G_t(t, J))] \end{aligned} \quad (7)$$

Discriminator and a pyramid densely-connected encoder-decoder improve the quality of the generated image by the network.

**GCA Net.** GCA Net adopts lightweight architecture, and used an end-to-end gated context aggregation network for dehazing.

The main contributions as fellow:

- (a) GCA Net utilizes smoothed dilated convolution to avoid grid artifacts, which refers to various forms of images that do not exist on the imaged object.
- (b) Hazy input image is first encoded into a feature map by the encoder part, and enhanced by aggregating more contextual information and fusing functionality of different levels without downsampling. In particular, it uses smooth expanded convolution [32] and additional gate subnets.

GCA Net decodes the enhanced feature map to original image space to get target haze residue (Eq. (13)) by utilizing smooth dilated convolutions and additional gated subnetworks. GCA Net used a new dilated convolutional layer, in which,  $f$  is the input variable,  $\omega$  represents regular convolution layer and  $k$  is kernel size, it can be demonstrated as Eq. (14) by adding a dilated filter on convolution layer.

$$(f \otimes_r w)(i) = \sum_{j=1}^k f[i + r * j]w[j] \quad (8)$$

When dilation rate,  $r = 1$ , dilated convolution will return to be regular convolution.

More, GCANet proposed extra gated fusion sub-network  $\mathbf{G}$  which consists of one convolutional layer with a kernel size of  $3 * 3$  to extract feature maps from different levels ( $\mathbf{F}_l, \mathbf{F}_m, \mathbf{F}_h$ ), and GCANet get three weights ( $\mathbf{M}_l, \mathbf{M}_m, \mathbf{M}_h$ ) from it. Combining the three feature maps and weights linearly, the total  $\mathbf{F}_o$  could be calculated.

$$(\mathbf{M}_l, \mathbf{M}_m, \mathbf{M}_h) = \mathbf{G}(\mathbf{F}_l, \mathbf{F}_m, \mathbf{F}_h), \mathbf{F}_o = \mathbf{M}_l * \mathbf{F}_l + \mathbf{M}_m * \mathbf{F}_m + \mathbf{M}_h * \mathbf{F}_h \quad (9)$$

The combined feature map  $\mathbf{F}_o$  will be fed to the decoder to obtain the selected haze residue. GCANet adopts simple mean square error loss used in previous papers.

$$\mathbf{r} = \mathbf{J} - \mathbf{I}, \hat{\mathbf{r}} = \text{GCANet} * (\mathbf{I}), L = ||\hat{\mathbf{r}} - \mathbf{r}||^2 \quad (10)$$

This is the residual between the haze-free image and the input blurred image,  $\mathbf{r}$  is ground truth and  $\hat{\mathbf{r}}$  is predicted haze residue. Using MSE loss as a loss function, the result can be improved by adding  $\mathbf{r}$  to the input haze image to obtain the final haze-free image.

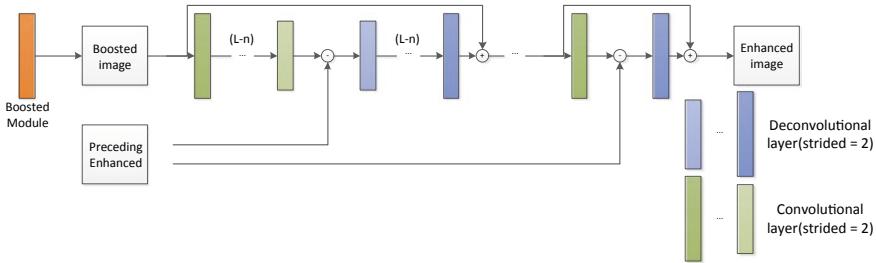
**Boosting algorithms-Multi-Scale Boosted Dehazing Network with Dense Feature Fusion.** MSBDN proposed U-Net architecture with dense feature fusion mechanism. In the encoding stage of the U-net structure, each layer feature is regarded as an input, and each layer feature map in the decoding stage is regarded as an output. Upsampling result of each layer in the decoding stage is regarded as the solution of the next iteration of the current feature map. The enhancement strategy commonly used in denoising stepwise solution is perfectly applied here. The enhancement strategy formula is unified, that is, the current input is added to the result of the previous output, and then used as the next input. The proposed modules:

The Strength-Operate-Subtract (SOS) strategy is used in the decoder as an image recovery module, and gradually refines the results by combining the previous results with the current input as input.

$$\hat{\mathbf{J}}^{n+1} = g(\mathbf{I} + \hat{\mathbf{J}}^n) - \hat{\mathbf{J}}^n \quad (11)$$

Directly using the splicing method for Dense Feature Fusion (DFF) is less effective. Back-projection technology is an effective approach to generating high-resolution content by minimizing the reconstruction error between high-resolution estimation results [33].

$$\hat{\mathbf{H}}_{t+1} = \hat{\mathbf{H}}_t + h(f(\hat{\mathbf{H}}_t) - \mathbf{L}_{ob}) \quad (12)$$



**Fig. 2** The architecture of DFF module in decoder

$H_t$  denotes the result of the  $t$ th iteration, and  $H_{t+1}$  denotes the result of  $t + 1$ th. At this time, it is hoped that the error of the  $t$ th time will be fed back to the  $t + 1$ th time.  $L_{ob}$  denotes a low-resolution image obtained through the f-downsampling operator. By performing the same down-sampling on the super result  $H_t$ , the difference between it and the down-sampling result of the real ground truth images could be generated and fed back through the function  $h$  (Fig. 2).

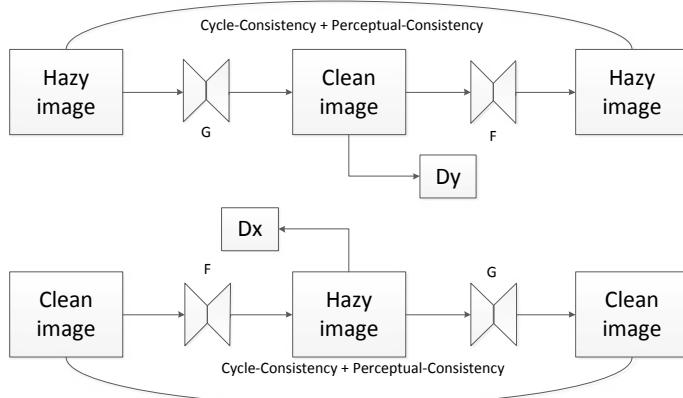
The method performs corresponding sampling processing on the previous result and error calculation with the current result, and obtains an error feedback through the function of the error which will be applied to the current result.

## 4.2 Generative Adversarial Network

After the original GAN [34] was proposed in 2014, it was widely used in various fields of artificial intelligence. GAN is composed of generator  $\mathbf{G}$  and discriminant  $\mathbf{D}$ , which are against each other and promote together to reach Nash equilibrium. Training is not accomplished until  $\mathbf{G}$  generates images close to ground truth images. There are numbers of variants and improvements to GAN, improving from its instability, difficult convergence and applicable filed, early variants such as used in denoise [35, 36], deblur [37] and derain [38] works.

**Cycle-Dehaze.** Cycle-Dehaze is based on CycleGAN [39] to make improvements, the former designed a cyclical counter-network learning image style migration, using unpaired images to train the network.

Cycle-Dehaze used end-to-end model for single image dehazing, and the proposed method trains the network by providing clean and haze images in unpaired ways instead of relying on the estimation of atmospheric scattering model parameters. Improving the quality of texture information recovery and producing better visually haze-free images. Integrating cycle-consistency and perceptual losses in view of the frame of CycleGAN. Due to GPU limitations, Cycle-Dehaze resizes input image to the size of  $256 \times 256$  pixel by bicubic downscaling. In order to improve the quality of input image, Cycle-Dehaze adopted Laplacian pyramid whose top layer has been changed by fundamental image to get high-resolution image.



**Fig. 3** Architecture of Cycle-Dehaze

Inspired by EnhanceNet [40], Cycle-Dehaze used perceptual loss to compare images in feature space instead of images in pixel space to strengthen image quality and metrics. Figure 3 shows two generators  $G$ ,  $F$ , and two discriminators  $D_x$  and  $D_y$ . Input image texture information will be saved and the haze-free will be output. Cyclic consistency loss takes higher weight since the color information may be lost if perceptual loss is higher.

Cycle-Dehaze employed feature extractor in VGG16 2nd and 5th layers. Cycle-Dehaze employed low-resolution images as input, so its input needs to be reduced to a preprocessing step, in order to reduce image distortion when resizing.

Cycle-Dehaze used the Laplacian pyramid to enlarge the low-resolution image instead of using bicubic method directly. The latter part of the paper shows experiments on datasets NTIRE 2018, distinctly,  $256 \times 256$  picture input limits its scope of application.

**FD-GAN.** FD-GAN [41] improves the performance of the algorithm from another perspective. They developed a fusion-discriminator to integrate frequency information as additional priors and constraints. This discriminator improves the performance of the generator's generated more realistic image from the side.

The HF (High frequency) component in the image represents those parts of image where the intensity changes rapidly, for example, sharp edges, textures, and fine details. On the contrary, the part of image whose intensity value changes slowly is the LF (Low frequency) area, such as the smooth area. After removing high frequency details, LF emphasizes the brightness, color and contrast information of the image, and it could make color comparison easier.

The mechanism of HF and LF is used to improve the performance of the discriminator, which in turn can act on generator to make it generate more realistic dehazed images.

Regular GAN network loss is used:

$$L = \alpha_1 L_1 + \alpha_2 L_S + \alpha_3 L_p + \alpha_4 L_G \quad (13)$$

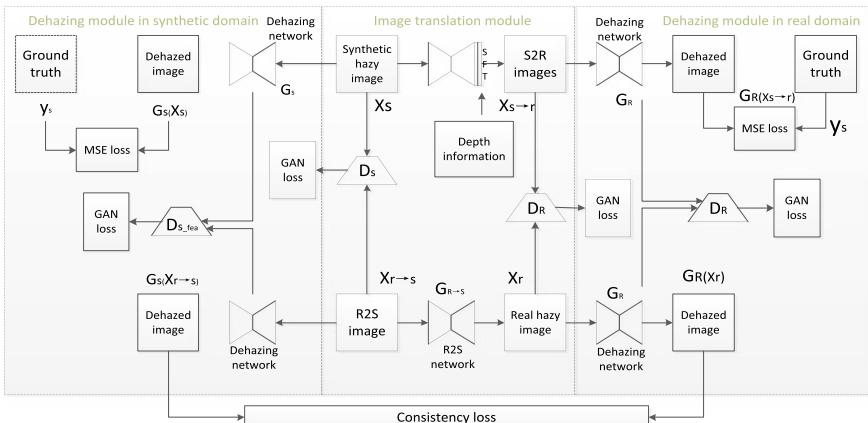
$L_1$  is pixel-wise loss and  $L_S$  is SSIM loss which are usually used in image processing,  $L_p$  represents perceptual loss and  $L_G$  stands for adversarial loss,  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  and  $\alpha_4$  are positive weights.

Another highlight in this paper is that they proposed a new synthetic dataset for training, which includes different indoor and outdoor images. FD-GAN proves that improving the quality of training data can strengthen the capabilities of the network.

The Encoder-decoder structure in generator can take full advantage of all extracted features maps from the shallow to the deep layers of the network, and FD-GAN chose nearest-neighbor interpolation for up-sampling due to the method will possibly reduce the artifacts generated by the image.

**Domain Adaptation for Image Dehazing.** Since the widespread problem in the dehazing algorithm, that is, the dehazing model trained only on synthetic data cannot be well extended to the processing of real haze images dehazing. DA proposed a novel end-to-end model for dehazing which effectively connect the synthetic and real-world hazy images, backbone of DA consists of three modules including dehazing module in synthetic domain, image translation module and dehazing module in real domain.

Two pictures denote the input of the image translation module as a composite picture and a real picture. The synthesized images are processed by the S2R module to generate a fake real world image, and the real image will also be processed by the R2S module to generate a fake synthesized image. The highlight is that the SFT can be added to the S2R module which requires a depth map. For SFT, the depth map obtained through the binocular map is used in the S2R network, because the haze is a layered image in the real image, and the distribution of the haze in the artificially formed synthetic image is uniform (Fig. 4).



**Fig. 4** Flow chart of algorithms DA

The depth map can be used to maximize the comprehensive synthetic image is converted into a real image, which can effectively combine the information relationship between the depth map and the feature map to generate a more realistic image. DA loss is defined as:

$$\begin{aligned} L = & L_{tran} + \lambda_m(L_{rm} + L_{sm}) + \lambda_d(L_{rd} + L_{sd}) \\ & + \lambda_t(L_{rt} + L_{st}) + \lambda_c L_c \end{aligned} \quad (14)$$

$\lambda$  denotes the four weight parameters,  $L_{tran}$  denotes composed of the confrontation loss of the two generators and its feature loss.  $L_{rm}$  denotes the loss of the predicted image and the clear image, and  $L_{sm}$  denotes the loss of the non-predicted image and the clear image.  $L_{sd}$  is the converted image for clear image comparison (dark channel),  $L_{rt}$  is the mutation loss and the L1 regularity before the predicted image JR Gradient,  $L_c$  is the consistency of the image after the dehazing network with the real haze network.

Transfer learning shows great effect in the field, method of dehazing in this area needs further research.

## 5 Performance Evaluation

### 5.1 Metrics

The objective evaluation of the image dehazing effect is to use some quantitative measurement methods to automatically evaluate the image quality, so as to obtain a parameter reflecting the quality or the degree of loss as the evaluation result. The commonly used PSNR (Peak Signal to Noise Ratio) is treated as objective standard for evaluating images. It is generally used for an engineering project between the background noise and maximum signal), SSIM (Structural Similarity Index Measure [42] is an index to evaluate the loss and distortion of the fused image. It consists of three parts: correlation loss, brightness and contrast distortion) and LPIPS.

### 5.2 Implementation

We tested the network on dataset RESIDE (Table 1) with Nvidia GTX 1660 SUP, testing part imports RGB images with size of  $640 \times 480 \times 3$ .

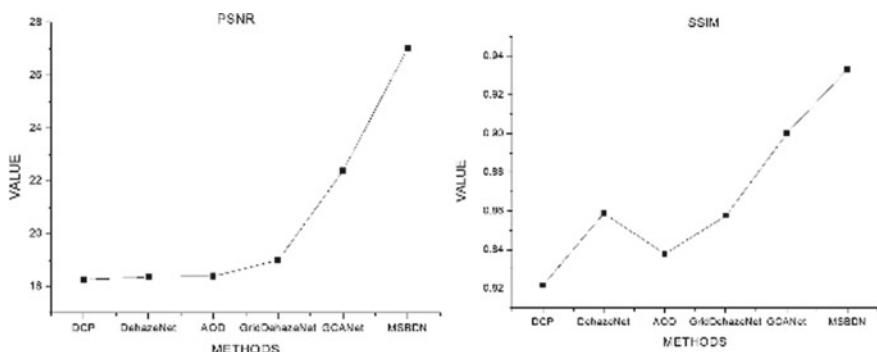
### 5.3 Results Evaluation

Compare the differences between the following results (Fig. 5 and 6; Table 2) and analyze them one by one. The problem studied in DCP is the image dehazing technology, which can use the haze density to estimate the distance of the object, and it can also restore the color and visibility of the image. These have essential applications in computer vision (such as three-dimensional reconstruction, object recognition). Researchers didn't know how to achieve this goal effectively before. In this paper, they found a statistical law and proposed an effective method of dehazing.

But the problem with the algorithm is that when the transmittance of the window is estimated through algorithms, there will be an edge effect of window size, and the estimated edge needs to be cut out, the step will extend image processing time. Processed image contrast is large, and some areas of image are not clean.



**Fig. 5** Results of different methods on synthetic datasets



**Fig. 6** Line chart of the results of the metric

**Table 2** Available datasets and information

Team	PSNR	SSIM	Average runtimes (s)
DCP	18.2701	0.8217	4.95
DehazeNet	18.3734	0.8590	1.19
AOD	18.4026	0.8380	0.41
GridDehazeNet	19.0212	0.8577	0.46
GCA-Net	22.4106	0.9001	0.64
MSBDN	27.0397	0.9333	0.77

As the most classic method among traditional algorithms, DCP is worthy of the public's high praise in terms of practicability and effectiveness. Later, He et al. improved it so that the algorithm reached the top effect of traditional methods.

As an early end-to-end network that used deep learning in the direction of dehazing, DehazeNet is based on manual features and goes beyond traditional methods. From manual to intelligent, it greatly improves the speed of image processing. It can be seen from the evaluation parameters that DehazeNet is much faster than traditional algorithms. AOD-Net, as a truly end-to-end lightweight model, unifies transmission map  $t(x)$  and  $A$  in Eq. (1) into one formula generates  $k(x)$  module, removes the step of calculating  $t(x)$ , and improves overall efficiency.

From the result picture below, we can see that AOD has a more saturated color than DehazeNet, and the latter darkens some areas in some images.

As the state-of-art method at the time, AOD achieved the effect of removing artifacts in a large area compared with the previous method, and also provided a direction for the subsequent dehazing method to remove image dirt.

The result of GridDehazeNet has a certain improvement over AOD-Net. However, as shown in the dehazing image, some uncovered image areas existed.

GCA-Net uses smooth dilated convolution instead of the original dilated convolution to solve the grid artifacts problem caused by the original dilated convolution, and through the gated fusion sub-network, the high-level and low-level features are integrated to improve the restoration effect and achieve a new height of dehazing effect.

MSBDN, as a better method now in the current open source algorithm, shows high results, which is a great improvement over the previous method. The back-projection technique can effectively fuse and extract features from different proportions for image dehazing. With the passage of time, new network architectures and modules have been proposed. Most of the problems in this field summarized above have been alleviated, which can be seen in terms of processing speed and image quality after restoration.

## 6 Conclusions

This paper summarizes traditional and recent deep learning dehazing methods. We select representative methods for experiments comparison, by showing and analyzing the changes and improvements of dehazing algorithms, two evaluation standards are used to measure. The evaluation criteria for images may no longer be limited to PSNR/SSIM.

The low-level image processing work including dehazing work would own clear but wrong details. Those images have worse PSNR results than blurry details, but its human vision the effect is better.

Most of the datasets in low-level image processing work are artificially synthesized. The network model obtained based on these data will have a domain gap when used in the real world. More use Real-world data is an improvement method, but the cost will increase accordingly. It is better to use limited real data and fine-tune the synthetic training model than mixed training.

The methods under the deep learning framework for low-level image processing still remains room for improvement, such as, in theories and algorithms, but for actual projects, the use of lightweight and could be adjusted in architecture or parameter model in simple structure may be easier to implement. Domain adaptation field problem, algorithm adaptation for specific scenes and projects may be a novel way to improve image quality and process speed more easily. The central idea of it is divergent from “no free lunch theorem”.

Low-level image processing is still in its infancy, the method based on the CNN network is still the mainstream in dehazing. The constant is that improving the dehazing effect by adapting the innovation of the modules in the network and the change of hyper parameters is the most direct way for a period of time. Simultaneously, for the GAN network, its unique image generation method and loss function also make it a common method in low-level image processing. Unpaired image training and image-to-image translation can be achieved by GAN.

All dehazing algorithms have a common target, which is to achieve great results that make people feel appropriate in different datasets and real hazy images, restore as much original image information as possible. **Solving the problems of domain adaptation. More appropriate and diverse datasets, innovative methods should be proposed.**

**Acknowledgements** Funding: This work was supported by the National Natural science Foundation of China (grant numbers 62006073).

## References

- Kaiming, H., Jian, S., Xiaoou, T.: Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(12), 2341–2353 (2011)
- Robby, T.T.: Visibility in bad weather from a single image. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Anchorage, AK, USA (2008)
- Sobhan, K.D., Mayukh, R., Debashis, S., Prabir, K.B.: Color cast dependent image dehazing via adaptive airlight refinement and non-linear color balancing. *IEEE Trans. Circuits Syst. Video Technol.* **36**, 2076–2081 (2021)
- Bolun, C., Xiangmin, X., Kui, J., Chunmei, Q., Dacheng, T.: Dehazenet: an end-to-end system for single image haze removal. *TIP* **25**(11), 05187–05198 (2016)
- Wenqi, R., et al.: Gated fusion network for single image dehazing. In: CVPR, pp. 3253–3261 (2018)
- Dongdong, C., et al.: Gated context aggregation network for image dehazing and deraining. In: WACV, pp. 1375–1383 (2019)
- Yossi, G., Assaf, S., Michal, I.: Double-DIP: unsupervised image decomposition via coupled deep-image-priors. In: CVPR, pp. 11026–11035 (2019)
- Xing, L., Masanori, S., Tankyuki, O.: Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In: CVPR, pp. 7007–7016 (2019)
- Deniz, E., Anil, G., Hazim, K.E.: Cycle-Dehaze: enhanced CycleGAN for single image dehazing. In: CVPR Workshops, pp. 825–833 (2018)
- Aditya, M., Harsh, S., Pratik, N., Murari, M.: HIDeGAN: a hyperspectral-guided image dehazing GAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 212–213 (2020)
- Bharath, R.N., Venkateswaran, N.: DehazeGAN: when image dehazing meets differential programming. In: WiSPNET 2020 Conference (2020)
- Wenhui, W., Anna, W., Qing, A., Chen, L., Jinglu, L.: AAGAN: enhanced single image dehazing with attention-to-attention generative adversarial network. *IEEE Access* **7**, 173485–173498 (2019)
- Nan, W., Yabin, Z., Fenglei, H., Haitao, Z., Yaojing, Z.: UWGAN: underwater GAN for real-world underwater color restoration and dehazing (2019). arXiv preprint [arXiv:1912.10269](https://arxiv.org/abs/1912.10269)
- Drews, P., do Nascimento, E., Moraes, F., Botelho, S., Campos, M.: Transmission estimation in underwater single images. In: ICCV Workshops, pp. 825–830 (2013)
- Zheng, L., Yafei, W., Xueyan, D., Zetian, M., Xianping, F.: Single underwater image enhancement by attenuation map guided color correction and detail preserved dehazing. *Neurocomputing* (2020)
- Yu, L., Robby, T.T., Michael, S.B.: Nighttime haze removal with glow and multiple light colors. In: ICCV, pp. 226–234 (2015)
- Jing, Z., Yang, C., Shuai, F., Yu, K., Chengwei, C.: Fast haze removal for nighttime image using maximum reflectance prior. In: CVPR, pp. 7418–7426 (2017)
- Jing, Z., Yang, C., Zha, Z., Dacheng, T.: Nighttime dehazing with a synthetic benchmark. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2355–2363 (2020)
- Sourya, D.D., Saikat, D.: Fast deep multi-patch hierarchical network for nonhomogeneous image dehazing. In: CVPR Workshops, pp. 482–483 (2020)
- Tiantong, G., Venkateswararao, C.I., Vishal, M.: Dense ‘123’ color enhancement dehazing network. In: CVPR Workshops (2019)
- Xu, Q., Zhilin, W., Yuanchao, B., Xiaodong, X., Huizhu, J.: FFA-Net: feature fusion attention network for single image dehazing. In: AAAI (2020)
- Howard, J.N.: Scattering phenomena (book reviews: Optics of the atmosphere: Scattering by Molecules and Particles). *Science* **196**, 1084–1085 (1977)
- Nathan, S., Derek, H., Pushmeet, K., Rob, F.: Indoor segmentation and support inference from RGBD images. In: ECCV’12 Proceedings of the 12th European conference on Computer Vision, pp. 746–760 (2012)

24. Jean-Philippe, T., Nicolas, H., Aurélien, C., Dominique, G., Halmaoui, H.: improved visibility of road scene images under heterogeneous fog. In: IEEE Intelligent Vehicles Symposium, pp. 478–485 (2010)
25. Cosmin, A., Codruta, O.A., De Christophe, V.: D-HAZY: a dataset to evaluate quantitatively dehazing algorithms. In: ICIP, pp. 2226–2230 (2016)
26. Cosmin, A., Codruta, O. A., Radu, T., De Christophe, V.: O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images. In: CVPR Workshops, pp. 754–762 (2018)
27. Cosmin, A., Codruta, O. A., Radu, T., De Christophe, V.: I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images. In: CVPR Workshops, pp. 620–631 (2018)
28. Boyi, W., Wenqi, R., Dengpan, F., et al.: Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* **28**(1), 492–505 (2019)
29. Shiyu, Z., Lin, Z., et al.: Evaluation of defogging: a real-world benchmark dataset, a new criterion and baselines. In: ICME, pp.1840–1845 (2019)
30. Boyi, L., Xiulian, P., Zhanyang, W., Jizheng, X., Dan, F.: AOD-Net: All-in-one dehazing network. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 1, pp. 4780–4788 (2017)
31. He, Z., Vishal, M.P.: Densely connected pyramid dehazing network. In: CVPR, pp. 3194–3203 (2018)
32. Zhengyang, W., Shuiwnag, J.: Smoothed dilated convolutions for improved dense prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2486–2495 (2018)
33. Michal, I., Shmuel, P.: Improving resolution by image registration. *Graphical Models and Image Processing*, pp. 231–239 (1991)
34. Ian, G., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
35. Jingwen, C., Jiawei, C., Hongyao, C., Ming, Y.: Image blind denoising with generative adversarial network based noise modeling. In: CVPR, pp. 3155–3164 (2018)
36. Dong-Wook, K., Jae-Ryun, C., Seung-Won, J.: GRDN: grouped residual dense network for real image denoising and GAN-based real-world noise modeling. In: CVPR Workshops (2019)
37. Orest, K., Volodymyr, B., Mykola, M., Dmytro, M., Jiri, M.: DeblurGAN: blind motion deblurring using conditional adversarial networks. In: CVPR, pp. 8183–8192 (2018)
38. Jinchuan, P., Xuesong, C., Li, Z., Qiuhan, Z., Yong, Z.: Removing rain based on a cycle generative adversarial network. In: ICIEA, pp. 621–626 (2018)
39. Jun-Yan, Z., Taesung, P., Phillip, I., Alexei, A. E.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV, pp. 2223–2232 (2017)
40. Mehdi, S.M.S., Bernhard, S., Michael, H.: EnhanceNet: single image super-resolution through automated texture synthesis. In: ICCV, pp. 4491–4500 (2017)
41. Yu, D., Yihao, L., He, Z., Shifeng, C., Yu, Q.: FD-GAN: generative adversarial networks with fusion-discriminator for single image dehazing. In: AAAI (2020)
42. Zhou, W., Alan, C.B., Hamid, R.S., Eero, P.S.: Image quality assessment: from error visibility to structural similarity. *TIP* **13**, 600–612 (2004)

# A Study on the Importance of Tactile Stimulation on Immersive Experience for Digital Communication



Jia Feng , Zhe Qian , Guoli Chen , and Wei Wang

**Abstract** In the post-pandemic era, telecommuting, online education, and digital entertainment have been widely proposed when there is a need for keeping physical distance. Analyses in digital communications literature indicate that individuals heavily rely on not only visual and auditory but also tactile stimulations to generate a more vivid digital communication. However, previous studies were somewhat focused on harnessing visual and auditory stimulations while the importance of tactile stimulation of digital experience has yet to receive academic attention. Given that, this paper examines the role of tactile stimulations to discern differences in individuals' immersive experiences between digital communications and face-to-face communications. First, literature studies analyze current immersive digital experience, people's digital communication expectations, and the effects of tactile recognition. Thereafter the research follows a research method based on content analysis and the research results indicate that touch-enabling technologies that omit tactile stimulations deliver poor communications.

**Keywords** Tactile stimulation · Immersive experience · Digital communications

## 1 Motivation

Einstein's Theory on Time and Space suggests that space is not static but relatively dynamic and multidimensional [1]. The interaction cyberspace is becoming a reality along with the development and application of 5G and digitalized technology. Communicating through online virtual or hybrid alternatives has become a new normal.

When physical and spatial obstacles are difficult to overcome, or physical distancing is strictly required, communicating through digital media is an ideal alternative. Moreover, machine learning and big data tend to generate objective results, while human judgments tend to be more subjective. Acknowledging this paradox,

---

J. Feng · Z. Qian · G. Chen · W. Wang ()

Department of Humanities, Art and Design, Zhejiang Gongshang University Hangzhou College of Commerce, 66 Huancheng South Rd., Tonglu County, Hangzhou 311300, China  
e-mail: [002030034@zjhzcc.edu.cn](mailto:002030034@zjhzcc.edu.cn)

principles of media and communications are therefore subject to rational consciousness: paying attention to the content of information and ignore the media; separate communication from the scene of personal relationship; praise rationality and reject irrationality; mask spatial and geographical elements; dismember the body into separate organs [2].

## 2 Previous Studies

### 2.1 *Expansion of Virtual Geographic Environment*

With the development of mobility in digital gadgets, applying digital screens in communications does not only limited to entertaining industries [3]. “Flow is the essence of the new media” [4], Zygmunt Bowman believes that digital media have had a human being married with a virtual world of high mobility.

Technically, such virtual geographic locations as the online office and online chatrooms can be built with spontaneity and creativity. The geographic location is spatialized, and the cyberspace and global village promise digital media users interactive communications: the media users could live and work on smartphones with a single click, and those who are empowered by digital technology could travel without any physical limitation. Therefore, the increasing availability of digital media that use immersive elements has created a more exciting online experience. In turn, digital screens’ stickiness increases.

### 2.2 *Flexible Working Hours and Dynamic Social Life*

As advanced technology allows media users to interact and communicate at any time without any biological or geographic limitations. Moreover, currently, people boast Internet access 24/7 and could be present anywhere without actual physical presence. This means that they do not have to commute to designated places for different occasions separately. For example, going to see a doctor does not mean merely going to the hospital, as remote medical treatment is available especially when there is a requirement for physical distance for the sake of health and safety; Connecting with family members or friends through social media allow people to have continuous contact throughout the day, and provide people with instant communication regardless of distance; Holding online conferences or outsourcing does not confine staff members in the office only. Therefore, the implementation of Cyber-Physical-Social Systems (CPSS) has enabled working hours and social life of media users to be more flexible and spontaneous.

The internet-based interaction has had human beings’ biological clock digitalized-media users are able to have unlimited access to different media platforms, and record

and replay digital content. On one hand, those who are empowered with smart digital media should be equipped with multitasking skills; On the other hand, people bring work home or deal with domestic problems at the workplace, the boundary between leisure and work has been blurred.

### ***2.3 Companionship and Social Currency***

Digital media have strong social attributes. As a carrier of digital information, digital media welcomes user-generated media content and has a low threshold for digital content uploading. Furthermore, digital media acquires only fragmented attention from the media users, this in turn allure media users to keep going back to different apps [5]. Therefore, how media users consume digital media content determines how they establish digital relationships, as the media users gain a feeling of companionship by spending time with digital screens [6].

The “Uses and Gratification Mode” perspective stated that media users generate media contents, establish the connection with others, share information, and consume media contents from other users [7]: First, media users turn to different digital platforms to achieve communication motives [8]. In the studies of compensation psychologies, statistics conducted from Parasocial relationships and television by Alan suggested that people can obtain a sense of fulfillment by watching live broadcasts, and those who do not have enough sense of belonging would turn to online social activities for comfort [9], such source of psychological activities comes from establishing a quasi-social relationship with streamers.

Second, the perspectives of communications of digital media users change as time goes by [10]. Individuals nowadays mobilize information from various online resources and engage with others, and it has never been easier like today for digital media users to engage in an open dialogue with almost anyone else. Moreover, the acceptance and interactions with other digital media users can generate social currency which in turn could have impacts on people’s real life. Therefore, media user’s social currency has become a key driver for communications [11]. One of the main reasons why people perceive the importance of social currency of digital media over interpersonal relationships is the widening physical distance among human beings. But the actual communication patterns do not validate social currency due to a possible reason that physical distance generates emotional distance and increases the cost of in-depth interactions among individuals [12].

### ***2.4 Immersive Digital Experience***

Studies from media and communications suggested that the “narcotizing dysfunction” of teletopia cultivates individuals to engage with the pseudo-environment

to forget about their physical presence [13]. Studies of immersive digital experience contain the two following aspects: First, digital screens provide a relatively private and closed-off two-sided communication; Second, median audiences have their visual and auditory sensory enlarged by directly interacting with digital screens without the mediated assistance of either a computer mouse nor keyboards [14]. For example, immersive media experiences such as virtual text drive, virtual fitting room, virtual mock examination integrate and mobilize individuals' visual and auditory senses. However, whether such immersive experience is an effective communication depends on the decoding and encoding literacy of both media communicators and immediate audiences. Furthermore, to engage with immersive experience ask for media audiences to work on their consensus to neglect interruptions from the real world [15]. Therefore, with the limitations of media technology, digital communications are not 100% immersive.

Moreover, physical conditions and movements of media users only transformed into a quantitative display through big data and machine learning. Such digitalized presentation of physical presence omits media users' social relations and political identities, and descendant "human" to a "digital individual" with only biological characteristics [16].

## 2.5 *Bringing Tactile Stimulation Back into Digital Communication*

In a virtuous community, media audiences become active communicators through different communication channels [17]. For example, people attend live broadcasts, participate in online shopping, and send bullet screen comments. However, E-commerce platforms have already heard the need for exploring a tactile sense of online shopping, and other digital platforms have also witnessed similar awaking.

Communication Research studies stressed the importance to congregate the body, media environment, and media platforms into effective communication. Paul Levinson's media humanization theory holds that the media is constantly evolving to meet human needs accordingly and build a digital environment more suitable to human communications. Moreover, the theory of embodied cognition emphasizes that an individual's mental state is strongly linked with its tactile experience. Peters has had his doubts stated in *Speaking into the Air*: to what extent could physical presence be absent in communication [18].

Digital technology empowers individuals to transcend the limitations and fragilities of the physical body, but digital media should also explore the potential of integrating tactile stimulation with media users [19]. Scholars who supported the absence of tactile presence put more value on communicating on spiritual levels [20], but effective communications require the combination of the body and the mind [21]. Amidst the wide application of VR, AI, and 5G into different industries,

realizing the positive impact of tactile stimulation in digital communication is of great importance.

To sum up, studying the effect of immersive experience as a means for communication has been widely applied in various fields. Most of the researches was conducted by studying how media users interact with digital media through hand clicking or body movement while few studies based on the role of tactile sentiment in an immersive digital environment for communication efficiency had been published. By studying the relevant literature on the development of media intelligence, media audience and immersive experience, and the relationship between physical presence and communication efficiency, this paper confirms the importance of tactile stimulation to digital communication. Previous theories and research results provide research theories, but in-depth studies on improving the immersive experience regarding tactile sensing were few. Therefore, this study analyzes the contributors that without experience disparities between online and offline communication, research results help bridge the gap between online and offline communications.

### 3 Research Methods

#### 3.1 Research Strategy

The findings suggest that the role of tactile stimulation is poorly addressed in digital communication, therefore there is a need to find out whether the absence of tactile stimulations decreases the depth of digital communication experience.

The content analysis method is employed in this paper to evaluate how digital communications were conducted when there is a need for physical distancing, the same criteria were applied to evaluate face-to-face communication to find out the role of tactile stimulation.

To ensure that the digital communications of different individuals, either negative or positive, were evenly experienced, collected literature is dated from January 2020 till now, as digital communication was universally applied since the outbreak of the COVID-19 pandemic. To increase the reliability of the study results, the author collects both Chinese and English literature from CNKI and [ScienceDirect.com](#): based on keywords “tactile stimulation”, “immersive experience”, and “digital communication”, 3,059 relevant literatures were found out (339 in Chinese and 2,720 in English), among which 1,428 were selected as the research samples.

#### 3.2 Research Criteria

To maintain the focus on how do digital communications differ from face-to-face communications, the following criteria were employed:

- (1) The location of the conducted communication
- (2) Dialogue aspects
- (3) Communication symbols that carry information
- (4) Interactivity of both parties when communicating
- (5) Information uncertainty.

## 4 Research Results

The specific results are presented in Tables 1, 2 and 3.

Table 1 shows that either digital conversations conducted through email, text messages, or video conferences, turn-taking is the only way to communicate, because interaction such as Internet breakdown, conversation overlaps, background noise would result in information misunderstood as voice or text is the only carrier of the information. Communications that rely on turn-taking are more time-consuming, considering there will be response delays, little timely feedback, or stilted conversation flow. Various surveys showed that people are more inclined to work at work sites.

**Table 1** The communication style of teleworkers and office workers

Factors	Telecommuting	Physical presence
Location	Apps	The office or other meeting places
Dialogue aspects	Turn-taking	Turn-taking, overlap
Communication symbol	Voice, context, text, pictures	Office environments, voice, text, eye contacts, facial expressions, body language
Interactivity	Poor	Strong
Information uncertainty	High	Low

**Table 2** The communication style of online education and offline education

Factors	Online class	Classroom in general
Location	Apps	Classroom, laboratory, training ground, etc.
Dialogue aspects	Cramming	Communicating
Communication symbol	Spoken discourse, written discourse, pictures	Study environment, spoken discourse, text, pictures, eye contacts, facial expressions, body language
Interactivity	Poor	Strong
Information uncertainty	High	Low

**Table 3** The communication style of digital entertainments and interpersonal communication

Factors	Digital entertainments	Interpersonal communication
Location	Apps	Classroom, laboratory, training ground, display ground
Dialogue aspects	Human-computer	People to people
Communication symbol	Digital screens	Surroundings, voice, eye contact, facial expression, body language, smell, temperature, light, and shadow effect
Interactivity	Poor	Strong
Information uncertainty	High	Low

While telecommuting reduces physical commuting and the work activities can be conducted through cloud-based collaboration and video conferencing, new communication problems emerge as well: in terms of communication between managers and employees, on the one hand, managers could not draw employees' profiles simply based on computer data, as employees' other qualities such as personalities, work proactivity, level of stress tolerance are also important communication skills but hard to evaluate through digital communications; on the other hand, communicating through digital screens lose the structure of work hierarchy, because social distance can be hardly sensed through digital screens; In terms of interactions among employees: physical presence enables them to establish bonding through dining, conducting group projects, etc., which would in turn increase trust among employees.

Table 2 demonstrates that during online education lecturers are the ones giving presentations while students are the ones cramming all the information within a limited time. As teaching is performed outside of the classroom, the lecturers could capture students' attention with voice and text only. Various surveys have shown that both the lecturers and students are more likely to attend classes traditionally.

In terms of communication between lecturers and students, at online classroom lecturers are senders while students are receivers. On the contrary, when the lecturers and students are in the classroom, they interact and communicate more freely from doing group discussions, group projects, and flip-over classrooms. Moreover, the physical presence of the lecturers sets a very good educational example by itself, as the saying goes "instruct and influence others by one's word and deed; in terms of communication among students, every classroom represents a small community where students watch and learn from each other.

As shown in Table 3, digital entertainments provide people with a digital companion of different forms. Spending time with a digital companion is cost-effective and bring individuals with the false impression that they are interacting with human beings. However, virtual companionship is the simulation of the real-life experience and expectation, and individuals communicate with their digital companions only through eye movements and finger scrolling. With today's technology limitations, emotion bonding through hugging, touching, and stroking that take place in

real life could not take place between humans and computers naturally, therefore the current digital companions could not replace human companions.

The reasons are as follows: first, although spending time with a digital companion reduce the level of social awkwardness, it also weakens the new experience brought by the uncertainty of interpersonal communication. Second, digital companion provides individuals with 24/7 telepresence, but such presence could only be detected through visual or audio stimulation. Furthermore, the absence of tactile stimulation would in turn entice individuals to look for visual and auditory excitements of high intensity to counterparts the void of tactile stimulations. Third, as parts of the social unit, individuals develop and maintain relationships with their families, friends, and loved ones mainly through physical contacts, therefore digital screens are never an ideal replacement for bonding.

## 5 Conclusion

Individuals bound through the diffusion of their intentions via tactile stimulation. As physical contact can be incorporated into meaning-making, and directly or indirectly influence how people perceive themselves, objects, and the physical world. On the contrary, deep learning-based interaction through digital screens has taken physical contacts out of the communication system, and those who struggle to keep up with digital communications or digitally excluded would disconnect either physically or emotionally. Moreover, it is even harder for those who are less privileged in Internet skills to conduct digital communications. Therefore, when individuals or social groups “meet” on digital media for the first time, there are fewer foundations for trust-building among different parties, and the higher level of uncertainties in communications the lower level of communication efficiency. This paper believes that digital communication works only when people concentrate and convince themselves that such an immersive digital environment exists. Such digital interaction is self-patronized and therefore could only be considered as a side product of face-to-face communication, not a long-term remedy for connection.

**Acknowledgements** This work was supported in parts by the Planning subject for Education and Science Programming of Zhejiang, China (2021SCG232) and the Research project for Philosophy and Social Science Programming of Hangzhou City (M21YD003).

## References

1. Hai-long, L.: What is the body problem from the perspective of communication. *News and Writing* **11**(1) (2020)
2. Wei, S.: The body of communicators: communication and the evolution of presence-consciousness subject, body-subject and intelligent subject. *International Press* **40**(12), 83–103 (2018)

3. Gui-wu, G., Qi, W.: Vertical video mode: a creative choice of video presentation in mobile scene. *Journalism Commun. Rev.* **72**(03), 8–107 (2019)
4. Lu, L.: On the quadruple influence of new media on urban space. *J. Sichuan Normal Univ. (Soc. Sci. Ed.)* **39**(04), 165–169 (2012)
5. Jia-yi, S.: Opportunities and challenges for the development of short video media. *China Television* **08**, 73–76 (2018)
6. Wei-wei, M.: Research on the Reconstruction of Social Relations in Live Webcasting. *J. Chongqing Univ. Posts Telecommun. (Soc. Sci. Ed.)* **6**, 96–102 (2018)
7. Berger, K., Klier, J., Klier, M.: A review of information systems research on online social networks. *Commun. Assoc. Inf. Syst.* **35**, 145–172 (2014)
8. Rubin, A.M.: *Uses-and-Gratifications Perspective on Media Effects*. Lawrence Erlbaum, pp. 525–548 (2002)
9. Hazan, C., Shaver, P.: Romantic love conceptualized as an attachment process. *J. Pers. Soc. Psychol.* **52**(3), 511–524 (1987)
10. Schiappa, E., Allen, M., Gregg, P.B.: Parasocial relationships and television: a meta-analysis of the effects. In: *Mass Media Effects research: Advances Through Meta-analysis*, pp. 301–314. Routledge (2003)
11. Bar-Anan, Y., Liberman, N., Trope, Y.: Automatic processing of psychological distance: evidence from a Stroop task. *J. Exp. Psychol. Gen.* **136**(4), 610–622 (2017)
12. Stephan, E., Liberman, N., Trope, Y.: Politeness and psychological distance: a construal level perspective. *J. Pers. Soc. Psychol.* **98**(2), 268–280 (2010)
13. Vascalou, A., Joinson, A. N. D.: Courvoisier. Cultural differences, experience with social networks and the nature of “true commitment” in Facebook. *Int. J. Hum.-Comput. Stud.* **8**(10), 719–728 (2016)
14. Xin, Z.: Analysis of aesthetic psychology of mobile phone vertical screen image from the perspective of immersion theory. *Northern Media Research* **06**, 76–80 (2020)
15. Si-ying, G.: On the application of immersive media in digital retail space. *Bus. Econ. Res.* **22**, 5–9 (2020)
16. Jia-chong, Q.: Investigation from the perspective of physical landscape and digital body-media technology in communication. *News and Writing* **11**, 20–27 (2020)
17. Yin, L., Hui-min, G.: Embodied interaction: an empirical interpretation of man-machine relationship in the era of intelligent communication. *News and Writing* **11**, 28–36 (2020)
18. Peters, J.D.: History of ideas about the spread of empty speech. Shanghai Translation Press (2017)
19. Deng, Z.: Body: an important dimension of the evolution of communication technology. *Young journalists* **33**, 29–30 (2020)
20. Hai-long, L.: Body agenda and the future of communication studies. *International Journalism* **40**(02), 37–46 (2018)
21. Zhuang, D.: Body: an important dimension of the evolution of communication technology. *Young Journalists* (33), 29–30 (2020)

# Single Shot Tooth Mark Detector for Tongue Diagnosis in Traditional Chinese Medicine



Xiaodong Huang and Li Zhuo

**Abstract** The Tooth mark is an important attribute of tongue diagnosis in Traditional Chinese Medicine (TCM). The recognition results obtained by current methods heavily rely on the results of tongue image segmentation. To solve the problem, we regarded the tooth marks recognition as a task of object detection and improved the original single shot detector (SSD) to detect the tooth marks. We removed the last two prediction layers of SSD and set the aspect ratios of the prior box to 1 based on the statistical data of the size and aspect ratios of tooth mark regions. Then we designed the multiple feature fusion module to combine the multi-scale features and embedded them hierarchically into the network to transfer the semantic information from deep layers to shallow layers. Furthermore, we also developed a feature enhancement module to improve the distinctiveness of features. The experimental results showed that the proposed method achieved 96.8% in terms of accuracy, which is significantly better than the current methods.

**Keywords** Tongue diagnosis · Traditional Chinese medicine · Object detection · Convolutional neural network

## 1 Introduction

Tongue diagnosis is one of the most important diagnostic methods in Traditional Chinese Medicine (TCM) and has been widely used for thousands of years [1]. Now, tongue diagnosis is still one of the most active approaches in the field of complementary medicine due to its advantage of simplicity, effectiveness, and non-invasion [2]. Tongue diagnosis is that the doctor observes the patient's tongue body and then assesses his health state [3]. Tongue diagnosis results heavily depend on the doctor's knowledge and experiences, and are also affected by the inspecting

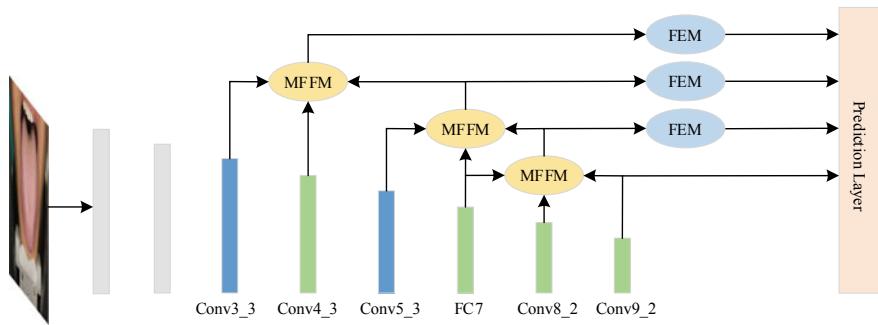
---

X. Huang · L. Zhuo

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China  
e-mail: [huangxiaodong@emails.bjut.edu.cn](mailto:huangxiaodong@emails.bjut.edu.cn)

X. Huang

Henan University of Science and Technology, Luoyang 471000, China



**Fig. 1** The architecture of TongueDet

circumstances. During the past several decades, researchers have developed many automated tongue diagnosis systems for the objective and quantitative diagnosis results [4–6].

During tongue diagnosis, tooth mark is an important attribute that has rich disease information. The tooth-marked tongue means there are some teeth marks on the tongue body. According to the theory of TCM, the tooth marks are related to the syndrome of deficiency of spleen qi. Because of spleen qi deficiency, the spleen cannot transport water and dampness, which causes the tongue to become fat. Then the tongue body compresses against the adjacent teeth, resulting in tooth-marks. These tooth marks mainly appear along the tongue's contour and vary hugely in color, shape, appearance, and size, which makes the automated recognition of tooth marks a challenge. Figure 1 illustrates the normal tongue and the tooth-marked tongue. Image (a) is a normal tongue image for reference. Image (b) is a tooth-marked tongue image whose contour of the tongue body is obvious curvature caused by tooth marks. Image (c) is a tooth-marked tongue image, but whose contour of the tongue body is non-obvious curvature.

Current methods of tooth mark recognition mainly utilize the curvature of the contour of the tongue body to get the suspected regions and then incorporate other handcrafted features to identify the tooth marks. However, even the most effective segmentation methods still cannot obtain the completely accurate tongue's contour, which could affect the recognition of tooth marks.

Furthermore, in some cases, the curvature of the contour of the tongue body is not obvious and even is smooth, which causes the existing methods invalid.

To solve the problem, we regarded the recognition of tooth marks as an object detection task and presented a novel single shot detector to detect the tooth marks, which directly outputs the tooth mark bounding boxes in a tongue image. The recognition results are robust to the results of tongue image segmentation and can provide more information about tooth marks, including the position, number, and even the degree of tooth marks. To our knowledge, the proposed detector is the first work that introduces the object detection network to recognize the tooth marks. We term this proposed detector as TongueDet.

## 2 Related Works

### 2.1 Recognition of Tooth Marks

The mainstream methods of tooth mark recognition usually acquire the suspicious zone based on the concavity of the contour of the tongue body and combine the other handcrafted features, such as color, brightness, the gradient of the concave region, to recognize the tooth marks. Wang [7] proposed a method to detect the tooth marks, which first calculated the slope of the margin of the tongue and the length and degree of the concave regions, and then used a threshold value to identify the teeth marks. Shao [8] exploited the features of concave and change of brightness and classified the tooth marks by thresholding feature values. Li [9] proposed a three-stage approach to recognize the tooth marks, including the proposal of suspected regions, extracting deep features by a convolutional neural network, and obtaining the results by a multiple instance classifier.

In summary, these methods heavily depend on the information of the indentations caused by tooth marks. So, the segmenting results of the tongue body can affect the results of recognition of tooth marks. Moreover, these methods cannot solve the inconstant performance of recognizing tooth marks when the tooth mark regions are not concave.

### 2.2 Object Detection Based on CNNs

After 2012, many new methods based on convolutional neural network (CNN) have kept emerging in the field of object detection and have gained remarkable success. These methods generally can be divided into two categories: the two-stage methods, and the one-stage methods.

The two-stage methods include two decoupled stages: proposal generation and box refinement. The region-based CNN (R-CNN) [10] is the most representative method, which is regarded as the first work of introducing CNN into the field of object detection. R-CNN is the foundation for many noted methods, including Spatial Pyramid Pooling (SPPNet) [11], Fast-RCNN, and Faster-RCNN. Though two-stage methods are quite effective, they are of low computational efficiency.

To improve the detection efficiency, one-stage methods directly predict the class probabilities and position coordinates of objects, removing the stage of proposal generation. YOLO [12] and SSD [13] are the two most famous one-stage methods for object detection. Based on YOLO and SSD, various detection methods are put forward to improve detection accuracy and speed.

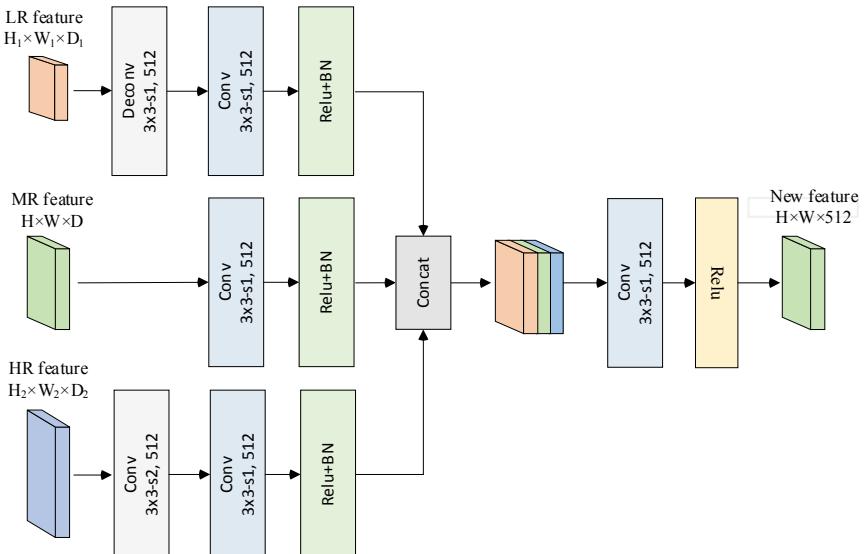
### 3 Methodology

#### 3.1 Architecture of TongueDet

Figure 2 shows the overall network architecture of TongueDet. TongueDet inherits the convolutional and box prediction components from original SSD, which employs the convolutional neural network to produce a fixed number of predicted bounding boxes and score the presence of different classes of objects in those boxes, and produce the final results by the non-maximum suppression. Compared with general object detection, the tooth mark detection has a simple scene and one category target. So, we modified SSD to adapt to the detection task of tooth marks.

We run K-means clustering on the training boxes of the tooth mark detection dataset to get the statistical information about the height, the width, and the aspect ratios of the bounding boxes of the tooth mark regions. We converged at five clusters and showed the result in Table 1.

Based on the results, it can be found that the aspect ratio of the most tooth mark regions is approximately equal to 1. Therefore, we retained the aspect ratio of the prior boxes to 1 while removing other settings in the original SSD. Moreover, we removed the last two layers of the original SSD under considering the size of the tooth mark regions and introduced two specific modules to improve the detector's performance.



**Fig. 2** The structure of multiple feature fusion module

**Table 1** The results of k-means

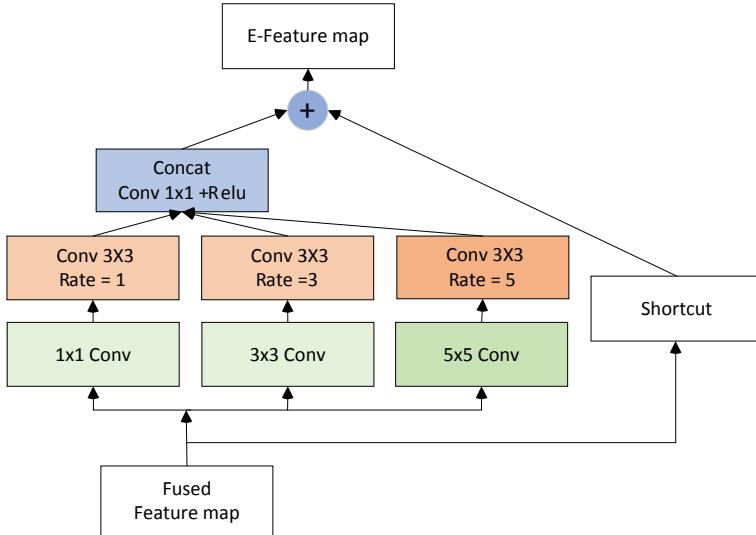
Width	Height	Aspect ratio
57	60	0.92
62	66	0.94
51	53	0.94
67	71	0.95
74	80	0.96

### 3.2 *Multiple Features Fusion Module (MFFM)*

Generally, CNNs perform consecutive pooling operations or convolution striding operations to learn increasingly abstract feature representations, and the deeper layers have more abstract semantic information. However, the information of the small objects is often weakened with the decrease in spatial resolution of the feature maps. Because of the small size of the tooth mark regions, the detector mainly detects the Conv4\_3 and FC7 to find the tooth marks. The two shallow layers usually have inadequate semantic information, which makes it difficult to recognize the tooth mark accurately. Therefore, we proposed a new multiple feature fusion module and embedded it iteratively into the backbone to enhance the classification capability of the detector, which can combine the semantic information from the deeper layers with the low-level vision information from the shallower layers. Figure 2 illustrates the structure of the multiple feature fusion module.

### 3.3 *Feature Enhancement Module (FEM)*

In some tongue images, the tooth marks are crowded, or the color of the tooth mark is similar to the color of the tongue body, which is a disadvantage to an accurate recognition of tooth marks. To solve the problem, we added a feature enhancement module (FEM) between the feature maps and prediction layers to improve the recognition of occluded tooth marks and to reduce the influence of similar colors from the tongue body. The FEM is inspired by the RFB [14] which includes a multi-branch convolutional block and a shortcut connection. The FEM simulates the human visual system and can generate more discriminative features, which is helpful for accurate object detection. The structure of the FEM is shown in Fig. 3.



**Fig. 3** The structure of features enhancement module

## 4 Experiments and Results

### 4.1 Dataset

We evaluated the performance of our detector on a dataset for tooth-marked tongue recognition, including 300 tooth-marked tongue images and 200 health tongue images. All tongue images come from the Department of Chinese Traditional Medicine of Beijing Xuanwu Hospital. We used 250 tooth-marked tongue images to train our detector and test it by using other 250 images, including 50 tooth-marked tongue images and 200 normal tongue images.

### 4.2 Implementation Details

**Data augmentation:** We adopt the same strategy of data augmentation used in SSD.

**Training parameters:** The proposed detector is trained by employ the stochastic gradient descent algorithm with a batch size of 8, a momentum of 0.9. We initialized the convolutional part using the pre-trained VGG16 [15] in ImageNet [16]. The initial learning rate is 0.01 and decayed to its 1/10 after 10, 000 iterations.

**Implementation:** All networks were trained and tested with PyTorch on the Windows system with the acceleration of one NVIDIA Titan RTX GPU card.

### 4.3 Evaluation Criteria

We utilized the three metrics to evaluate the cognition results: accuracy, sensitivity and specificity, involved four variables: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

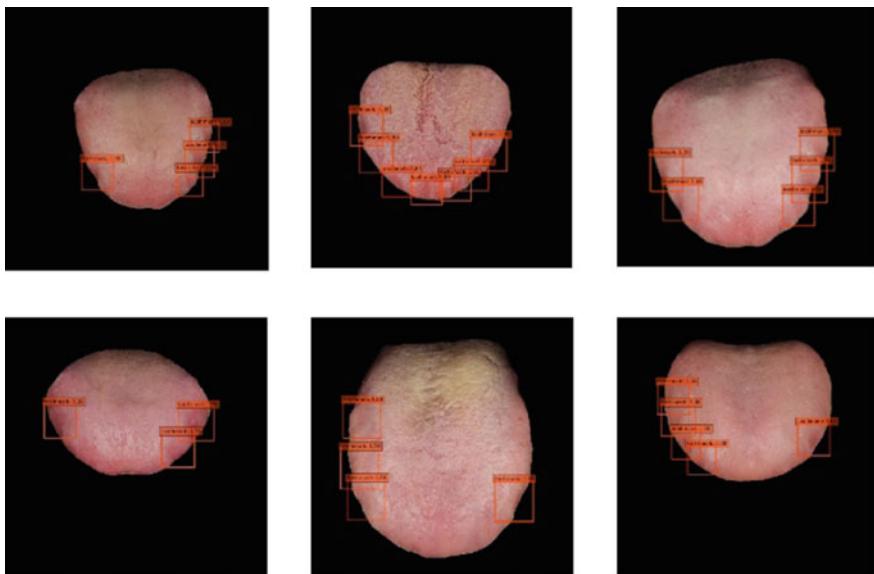
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

### 4.4 Experiment Results

- (1) Detection results of tooth marks in the tooth-marked tongue images. We tested the tooth mark detector on 50 tooth-marked tongue images. Figure 4 displays the detection results of six tooth-marked tongue images. It can be seen from Fig. 4 that our detector can recognize the tooth marks whenever the contours of



**Fig. 4** The detection results of tooth marks of tooth-marked tongue images

**Table 2** The comparison results of different recognition methods of tooth-marked tongue

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
Wang's [7]	49.8	97.4	86.8
Shao's [8]	60.4	42.7	76.6
Li's [9]	73.1	72.7	73.9
Our	96.7	96.0	97.0

the tongue body are curvaceous or not. The recognition results are not affected by the segmentation results of the tongue body.

- (2) Comparison with the existing recognition methods of teeth-marked tongue. We conducted a comparison with other existing recognition methods of the teeth-marked tongue on the test dataset, including 50 tooth-marked tongue images and 200 non-tooth-marked tongue images. Our tooth mark detector is an end-to-end object detection network based on convolutional neural network. The detector is powerful in feature extraction and representation, which applies MFFM to aggregate multi-scale features, and also employs FEM to further improve the discrimination of fused features. Benefited from the elaborated design, the tooth mark detector can achieve higher accuracy in recognition of tooth-marked tongue compared with other methods. The comparison results are itemized in Table 2.

## 5 Conclusions

In this study, we proposed a new method of tooth-marked tongue recognition in TCM. Different from the traditional approaches, our method employed the technology of object detection based on deep learning and can directly output the bounding boxes of tooth mark regions. We modified the original SSD to adopt the task of tooth marks recognition. And we removed the last two layers of the original SSD and set the aspect ratios of the prior box to 1, based on the statistical results of the size and aspect ratios of the tooth mark regions. Meanwhile, we designed a multiple feature fusion module to combine the high-level semantic information and low-level vision information. Furthermore, we inserted a feature enhancement module before every prediction layer to improve the distinctiveness of feature maps. Our tooth mark detector resolves the problem that the results of tooth mark recognition are affected by the results of tongue image segmentation and makes a great improvement in recognition accurateness compared with other methods.

In the future, there will be much work to do. One is to improve the performance of our model by using more tongue images to train it. Another is to tackle the problem of domain adaptation so that our model can be used to detect tongue images captured by different devices under different circumstances.

**Acknowledgements** This work is supported by National Natural Science Foundation of China [61871006].

## References

1. David, Z., Hongzhi, Z., Bob, Z.: *Tongue Image Analysis*. Springer, Singapore (2017)
2. Yuqi, L., Wang, Y., Nannan, S., Xuejie, H., Aiping, L.: Current situation of International Organization for Standardization/Technical Committee 249 international standards of traditional Chinese medicine. *Chin. J. Integr. Med.* **23**(5), 376–380 (2016)
3. Ziyin, S.: Basic theory of traditional Chinese medicine. *Chin. J. Integr. Traditional Western Med.* **17**(11), 643–644 (1997)
4. Chuangchien, C.: A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue. *Comput. Methods Programs Biomed.* **61**(2), 77–89 (2000)
5. Oji, T., Namiki, T., Nakaguchi, T., Ueda, K., Takeda, K., Nakamura, M., Okamoto, H., Hirasaki, Y.: Study of factors involved in tongue color diagnosis by Kampo medical practitioners using the Farnsworth-Munsell 100 hue test and tongue color images. *Evidence-Based Complementary Alternative Medicine* **2014**(3) (2014)
6. Yonggang, W., Yue, Z., Jie, Y., Qing, X.: An image analysis system for tongue diagnosis in traditional Chinese medicine. In: *International Conference on Computational and Information Science*, pp. 1181–1186. Springer (2004)
7. Dongxue, W., Hongzhi, Z., Jianfeng, L., Yanlai, L., David, Z.: A Novel feature extraction method of toothprint on tongue in traditional Chinese medicine. In: *International Conference on Information and Management Engineering*, pp. 297–305. Springer (2011)
8. Qing, S., Xiaoqiang, L., Zhicheng, F.: Recognition of teeth-marked tongue based on gradient of concave region. In: *2014 International Conference on Audio, Language and Image Processing*, pp. 968–972. IEEE (2014)
9. Li, X., Yin, Z., Qing, C., Xiaoming, Y., Zhang, Y.: Tooth-marked tongue recognition using multiple instance learning and CNN features. *IEEE Trans. Cybern.* **49**(2), 380–387 (2018)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587. IEEE (2014)
11. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788. IEEE (2016)
13. Wei, L., Dragomir, A., Dumitru, E., Christian, S., Scott, R., Chengyang, F., Alexander, B.: SSD: Single shot multibox detector. In: *European Conference on Computer Vision*, pp. 21–37. Springer (2016)
14. Songtao, L., Di, H.: Receptive field block net for accurate and fast object detection. In: *European Conference on Computer Vision (ECCV)*, pp. 385–400 (2018)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations* (2015)
16. Deng, J., Dong, W., Socher, R., Lijia, L., Kai, L., Feifei, L.: Imagenet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE (2009)

# Based on Machine Vision, Automatic Measuring System for Adhesive Coating of Caliper Tool



Xiaofei Wang , Xiaolei Zhang , Jing Wang , and Hua Fan

**Abstract** At present, it is very subjective to judge whether the amount of glue dispensing meets the requirements in the detection of glue coating, so there is a large error. Therefore, an automatic detection system of the caliper tool for glue coating based on machine vision is proposed. The system first by CCD cameras for electronic image acquisition, and then by using image processing algorithm for acquisition of image analysis processing automatic positioning to the target to be detected, then the automatic assignment of caliper, the target area according to the caliper area measurement location glue on the edge of the width, with a given range of standard compare whether meet the requirements. The experimental results show that the system has high accuracy, controllable measurement accuracy and short measurement time.

**Keywords** Adhesive testing · Machine vision · Caliper tool · Accuracy of measurement

## 1 Introduction

In the process of industrial testing, the demand for the quantity of gluing testing continues to increase. In order to ensure the processing quality and production efficiency, there is an urgent need for high-precision measurement of related technologies and instruments. At present, the vernier caliper measurement method can be used for the small width of the glue, but the measurement process is relatively complicated, and the measurement position is not only many but also has great subjective factors, and the measurement accuracy has a certain limit. Because the width of the glue is small in the actual industrial application, it is difficult to meet the requirements of the measurement speed and precision by using the traditional measurement method. Because the machine vision has the characteristics of a high precision and high degree of automation, this paper uses the machine vision method to measure the width of

---

X. Wang · X. Zhang · J. Wang · H. Fan  
Laser Institute, Shandong Academy of Sciences, Jinan 250000, China  
e-mail: [xiaoleizh@sdlaser.cn](mailto:xiaoleizh@sdlaser.cn)

the glue. Machine vision adopts non-contact measurement [1], which can not only detect the workpiece in a complex environment, but also detect the workpiece with high accuracy requirements. Because machine vision inspection has the advantages of fast speed, simple structure, high measurement accuracy and simple operation, it has become the mainstream detection tool of intelligent production line. In this system, machine vision technology is used to measure the width of the adhesive, and the industrial camera is calibrated with the Zhang Zhengyou calibration method [2]. The fusion image processing technology of BLOB analysis [3] and skeleton extraction [4] is adopted. Finally, the improved caliper method is used to extract the width of the adhesive to be measured [5]. The repeated tests show that the machine vision measurement of the gluing track width has a low misjudgment rate, fast beat and reliable operation. The design of the system effectively improves the gluing quality and safety performance.

In this experiment, machine vision method is used to measure the width of the glue. The measuring caliper method written by ourselves is used to measure the width of the glue. Finally, the user interface is designed through VS. Good experimental results are obtained by the system, which provides a basis for industrial testing.

## 2 System Scheme Design

After selecting appropriate hardware and equipping the system, adjust the camera, shooting plane and light source position to ensure that the image obtained is clear, with fewer excessive edges and strong contrast. Image segmentation algorithm is used to obtain the area to be detected, for the area to be detected automatically allocate the corresponding number of calipers, for each caliper sub-pixel precision edge search, obtain the width of multiple dispensing positions, and compared with the given dispensing width standard, draw a conclusion.

## 3 The Hardware System

The hardware system designed in this paper includes industrial camera, telecentric lens and annular light source. The selection of an appropriate hardware system determines the quality of the image collected, thus affecting the measurement accuracy.

### 3.1 Industrial Camera

The resolution of industrial cameras is related to the pixel size of the optical sensor and the optical magnification of the matched lens. The magnification of the lens is

$\alpha$ , the size of the image obtained by CCD is  $A \times B$ , and the resolution of CCD is  $M \times N$ , so the detail resolution of the camera on the photographed object can be calculated as follows:

$$\begin{cases} R_x = \frac{1/(M/a)}{\alpha} \\ R_y = \frac{1/(N/b)}{\alpha} \end{cases} \quad (1)$$

$R_x, R_y$  respectively are the minimum distance that the camera can distinguish between the horizontal and vertical images.

### 3.2 The Lens

In the case of selecting a lens, the working distance of the camera, the field of view of the lens, the focal length, the depth of field, and the distortion rate and spectral characteristics of the lens should be considered. In general, electronic components have a certain depth, so the use of telecentric lens can eliminate the focusing difficulty caused by ordinary FA lens, which leads to inaccurate measurement. At the same time, telecentric lens can also reduce the lens distortion caused by imaging problems.

### 3.3 The Light Source

The selection of light source will affect the information of the target to be measured, and the appropriate choice of light source will make the target to be measured clear. The subsequent image processing algorithm is relatively simple. If the light source is not suitable, it will increase the difficulty of image processing, and even affect the accuracy and speed of measurement. The annular light source can eliminate the factors caused by the poor reflection of the solder image and improve the imaging effect.

In this paper, according to the actual situation, choose a 5 million pixel industrial camera based on Ethernet protocol; High precision double telecentric lens with 0.113 magnification was selected. Ring light source is selected as the light source.

## 4 The Software System

### 4.1 Camera Calibration

In this paper, the relationship between system image coordinates and world coordinates is obtained by using the method of checkerboard calibration. The checkerboard was selected as  $7 * 7$  center calibration plate, and the size of each center of the circle was 56 mm. The internal and external parameters of the camera were obtained by camera calibration algorithm (Fig. 1).

The relationship between ideal coordinates and camera coordinates is as follows:

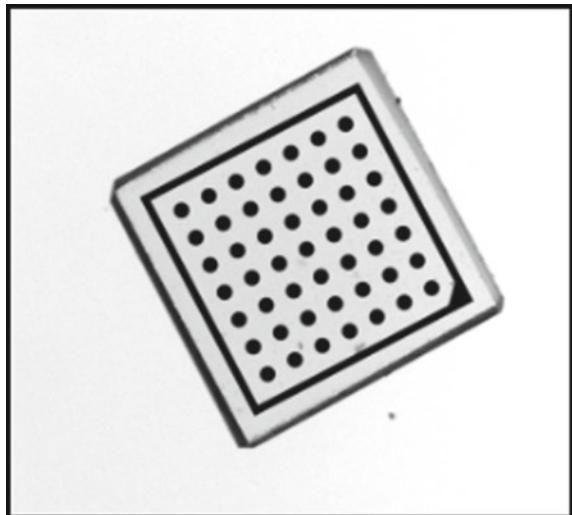
$$\frac{x}{f} = \frac{x_a}{z_a}, \frac{y}{f} = \frac{y_a}{z_a} \quad (2)$$

The mapping matrix between the world coordinate system and the ideal coordinate system is as follows:

$$z_a \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M \times R \times T \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = M \times N \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (3)$$

where, M is the internal parameter matrix of the camera; N is the camera external parameter; R is the relative model camera rotation; T is the run. M, R and T are expressed by a matrix as follows:

**Fig. 1** Checkerboard picture



$$\begin{aligned}
 M &= \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \\
 R &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = [a_1 \ a_2 \ a_3], \\
 T &= \begin{bmatrix} a_{14} \\ a_{24} \\ a_{34} \end{bmatrix}
 \end{aligned} \tag{4}$$

$$z_a \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M [a_1 \ a_2 \ T] \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{5}$$

Point P is associated with the homography matrix  $N = M[a_1 \ a_2 \ T]$  of point P in the matching template. Assuming that the matrix N is the internal and external parameters of the camera, the perspective mathematical model and the expressions of U and V after transformation can be obtained through changes, as shown in Eq. (6).

$$\begin{cases} u = \frac{f_x(a_{11}x + a_{12}y + a_{14})}{a_{31}x + a_{32}y + a_{34}} + u_0 \\ v = \frac{f_y(a_{21}x + a_{22}y + a_{24})}{a_{31}x + a_{32}y + a_{34}} + v_0 \end{cases} \tag{6}$$

Due to the external influence of the camera, the actual image points will have distortion, so the existence of distortion should be considered during reconstruction. Equation (7) can be obtained.

$$\begin{cases} x_u = \frac{a_{11}x + a_{12}y + a_{14}}{a_{31}x + a_{32}y + a_{34}} + u_0 \\ y_u = \frac{a_{21}x + a_{22}y + a_{24}}{a_{31}x + a_{32}y + a_{34}} + v_0 \end{cases} \tag{7}$$

Influenced by other factors, the camera will be distorted, and the expression leading to the offset of the image points is as follows:

$$\begin{cases} (x_d - x_u) = k_1 x_u (x_u^2 + y_u^2) + k_2 x_u (x_u^2 + y_u^2)^2 \\ (y_d - y_u) = k_1 y_u (x_u^2 + y_u^2) + k_2 y_u (x_u^2 + y_u^2)^2 \end{cases} \tag{8}$$

Internal parameter M can be calculated from Eqs. (7) and (8).

## 4.2 Positioning Module

The module uses BLOB analysis to obtain the required area, and then extracts the skeleton from the area, and automatically generates the required calipers through the extracted skeleton.

**Threshold segmentation.** In this paper, the maximum inter-class variance method is used to determine the adaptive threshold and automatically segment the target image. According to the gray histogram of the image, the target and the background of the image are distinguished. Theoretically, the larger the class variance between the target and the background, the greater the difference between the two parts of the image. When the target and the background are misclassified, the difference between the two parts will become smaller.

If the obtained segmentation threshold is  $T$ , the proportion of the number of target pixels to the total number of image pixels is  $W_0$ , and the average gray level is  $U_0$ . The ratio of the number of background pixels to the total number of images is  $w_1$ , the average gray level is  $u_1$ , the average gray level of the image is  $u$ , and the variance  $g$  between the target and the background image can be obtained as follows:

$$g = w_0 \times w_1 \times (u_0 - u_1)^2 \quad (9)$$

**Feature selection.** The simplest regional feature is the area of a region:

$$a = |R| = \sum_{(r,c \in R)} 1 = \sum_{i=1}^n ce_i - cs_i + 1 \quad (10)$$

To put it simply, for a binary image, the area of the region is the sum of the number of points in the region.

**Measuring module.** The steps to get the edge with caliper tool are as follows:

The edge of sub-pixel is accurately extracted by caliper method. Firstly, the gray value of profile line in the measurement area is obtained. The average gray value perpendicular to the profile line is used to represent the gray value of this point on the profile line. Then the profile line is smoothed to eliminate interference, and the curve is smoothed by Gaussian filtering. Then take the derivative of the smoothed profile line, and the extreme point is the position of the edge point. The edge with small gradient can be removed according to the threshold value. In this way, the corresponding edge degree can be obtained. The specific process is shown in Figs. 2, 3, 4 and 5.

In this paper, the caliper tool is generated equidistance on the extracted skeleton line contour, and then the method of measuring caliper tool is used to detect the edge pair width. Obviously, the more the number of caliper tools, the more accurate, and at the same time, the detection time is also relatively increased. In order to give consideration to both accuracy and efficiency, this paper uses the method of generating card scale at intervals, that is, a caliper is generated every 10 points, and

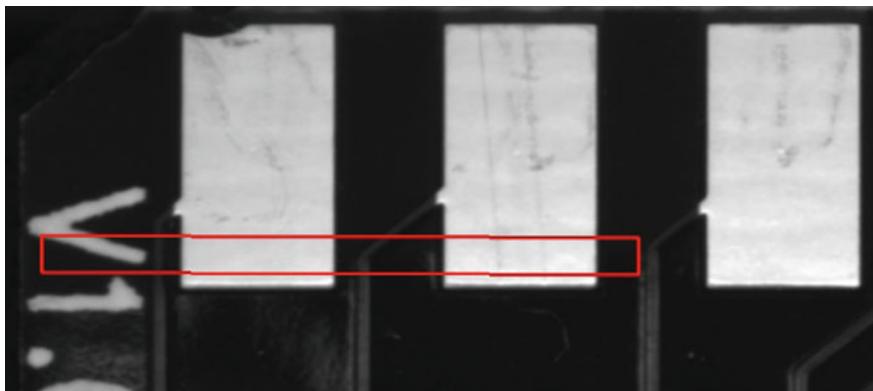


Fig. 2 Raw image and ROI region settings

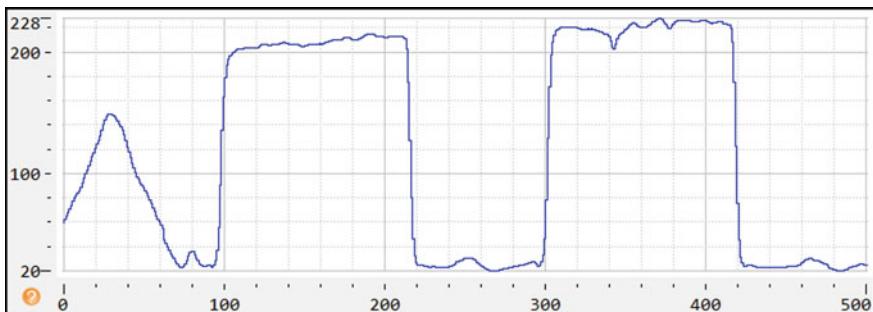


Fig. 3 Grayscale curve

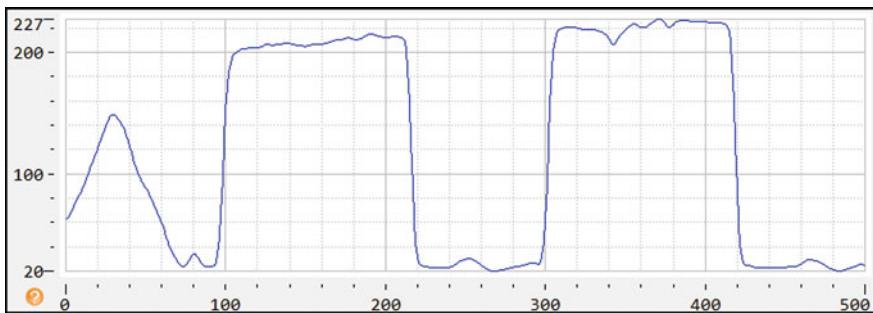
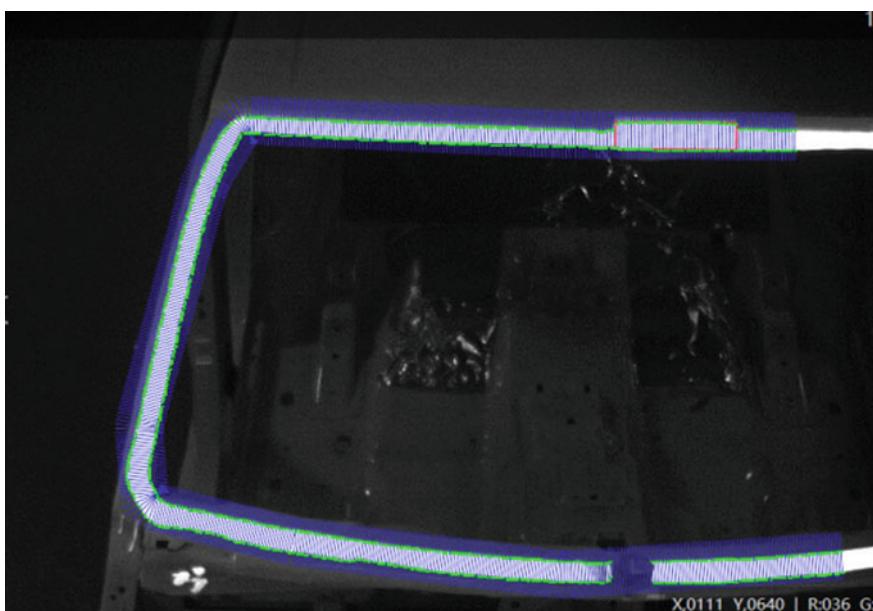


Fig. 4 Smooth curve



**Fig. 5** Derivative curve



**Fig. 6** Results the image

the width of the caliper is set as 5. The contour obtained by this algorithm in this paper is a sub-pixel edge (Fig. 6).

## 5 The Error Analysis

In order to observe the detection results, VS2015 software was used as the experimental platform. The ideal gluing track is 5 mm width for manual measurement

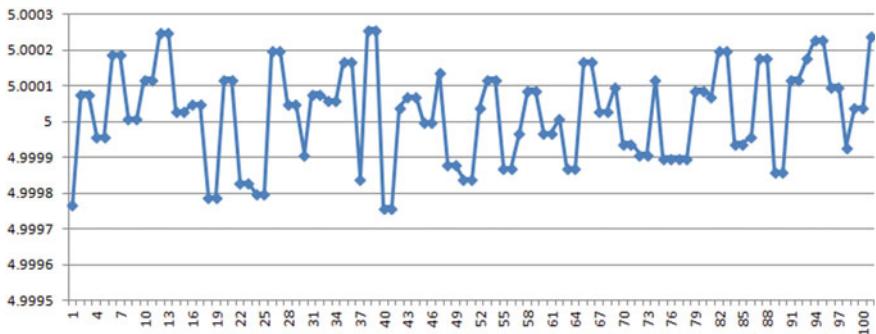


Fig. 7 The standard value of measurement

and machine vision measurement for 100 times respectively. The measured values are shown in Fig. 7. The experimental data show that the error value of the machine vision measurement is small compared with the ideal measurement value, and the measured values are within the required range of  $5 \pm 1$  mm with high detection accuracy.

Through the comparison of the above data and results, it can be seen that using machine vision to measure the width of the gluing track has little error and low misjudgment rate, which effectively solves the problems of the gluing position leakage and gluing leakage in the welding and assembly production line. From the time analysis, it can be seen that the manual measurement time is at least 180 s, and all the detection can not meet the production demand, but it can be completed with machine vision measurement 5 s, and the production can be met no matter how to improve the production pace due to the production demand in the later stage. It can be seen that the machine vision measurement technology is beneficial to improve the speed and detection accuracy of the coating position in the production line (Fig. 8).

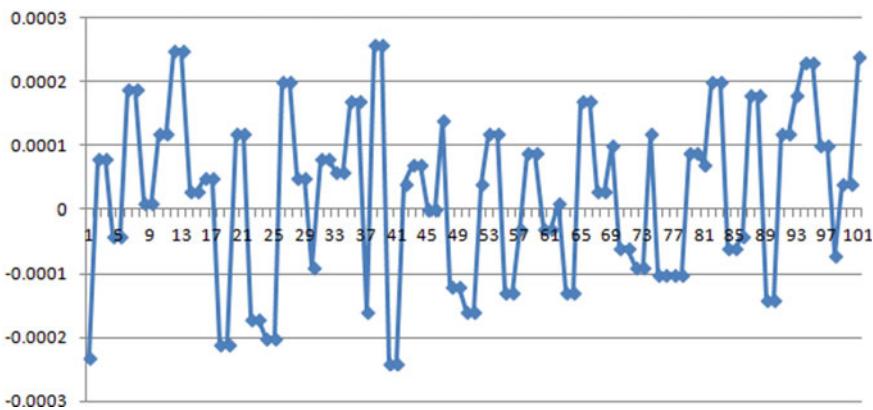


Fig. 8 Measurement error deviation

## 6 Conclusions

The system uses caliper tool to detect the sub-pixel edge point, and then carries on the width measurement. From the measurement results, it can be seen that the system solves the problem of measuring the glue-coating area manually and inaccurately. In addition, the system also has the characteristics of relative stability, strong robustness and fast measurement speed. The vision technology is used to measure the width of the gluing track, which realizes the intelligent measurement of the fluid, improves the working efficiency of the production line, and avoids the quality risk of glue leakage during assembly caused by the wide gluing track and the side assembly leakage caused by the narrow gluing track.

**Acknowledgements** This work was supported by Shandong Key Research Project (No.2019RKC01001).

## References

1. Zhuangwen, H., Huacai, L., Wengen, G.: Design and research of side coating detection system based on machine vision. *J. Anhui Eng. Univ.* **35**(5), 47–52 (2019)
2. Xiaojun, T., Zhi, Y., Jun, L.: An improved method for stereo camera calibration. *Acta Geodaetica et Cartographica Sinica* **35**(2), 138–142 (2006)
3. Fuzhong, W., Kaikai, Y.: A local threshold segmentation algorithm based on median filtering. *Electron. Meas. Technol.*, 162–166 (2017)
4. Shiquan, A., Meng, X., Qizhong, R.: School of Computer Science and Technology, Chongqing University of Posts and Telecommunications. *Comput. Eng. Des. Nat. Sci. Ed.*, 3170–3175 (2018)
5. Lushen, W., Jumin, X., Yun, H.: Machine vision based caliper tool nut real-time detection system. *Instrum. Technol. Sensor*, 50–55 (2020)

# Infrared Image Data Augmentation Based on Improved Image-to-Image Translation Network



Zizhuang Song , Jiawei Yang , Dongfang Zhang , and Yue Zhang

**Abstract** In practical application, deep learning method is often limited by the lack of training data and cannot achieve the expected results. In this paper, an infrared image data augmentation method based on improved image-to-image translation network is proposed. Firstly, the two-way feature fusion makes the generating network be able to extract more detailed features from the image, and feature-level alignment is used to ensure the semantic consistency between the translated images, which makes up for the defect of pixel-level alignment. Meanwhile, the self-attention mechanism is used to make the network pay attention to the foreground and background information at the same time. The above methods are applied to the low altitude sea surface visible and infrared image dataset, and results show that the proposed method can generate higher quality infrared images than CycleGAN and reduce the FID score by 15.78. Furthermore, by translating the visible image into the infrared image, the infrared image dataset is expanded, and the infrared object detection accuracy is improved. The of the infrared object detection model is increased by 5.28%.

**Keywords** Image-to-image translation · Data augmentation · Infrared image

## 1 Introduction

For the task of infrared target detection on the sea, the information received by infrared detector is always changing dynamically, which leads to the lack of uncertainty generalization ability of deep learning network for target features. The data-driven deep learning algorithm cannot achieve the desired performance in application scene with limited data. How to generate and expand the limited data to make the deep learning network achieve the desired performance becomes particularly important.

Compared with visible images, infrared images are less in quantity and sometimes inconvenient in data acquisition. When training object detection network, less infrared image data is easy to lead to over fitting, which affects the generalization

---

Z. Song ( ) · J. Yang · D. Zhang · Y. Zhang  
Beijing Institute of Remote Sensing Equipment, Beijing 100854, China

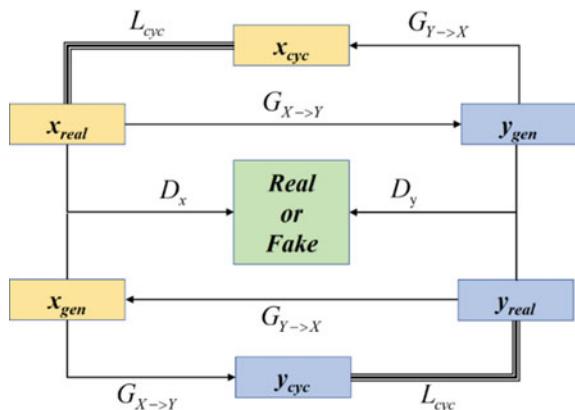
ability of the model. Li et al. have transformed visible ImageNet dataset into infrared image data by using GAN [1] and classified them, which achieved 91.28% accuracy and solved the problem of less infrared images [2]. Wang et al. have used the variational auto-encoder to eliminate the differences between modalities, and used the generated infrared and visible images to form 4-channel data to carry out cross-modality pedestrian detection [3]. Wang et al. have added joint pixel feature alignment to the GAN, generating infrared image by using visible image and carrying out cross-modality object detection [4]. Chen et al. have proposed a method of generating infrared images based on GAN to carry out infrared image data augmentation [5]. Hu et al. have proposed an infrared and visible face image translation network based on CycleGAN [6], improving network structure and designing a new loss function, which effectively overcomes the modality differences of images caused by different spectral characteristics [7].

The translation between infrared and visible image based on CycleGAN is a conversion of image modality, which retains the original image foreground–background information. At present, most of the image generation methods based on CycleGAN are pixel-level alignment without feature-level alignment, which is lack of semantic restrictions for the generated image.

## 2 Theory of the CycleGAN

CycleGAN uses two GANs to form a dual structure so that images between different modalities can be translated circularly, and the accuracy of image translation is ensured by the cycle consistency loss (Fig. 1). CycleGAN is an unsupervised network and the dataset does not require paired images. Taking the translation between infrared image and visible image as an example,  $X$  is the visible domain,  $G_{X \rightarrow Y}$  is the generation network from visible image to infrared image, and  $D_Y$  is the infrared domain discriminator to identify the authenticity of the infrared image translated by

**Fig. 1** CycleGAN network structure



visible image and the real infrared image. Similarly,  $Y$  is the infrared domain,  $G_{Y \rightarrow X}$  is the generation network from infrared image to visible image,  $D_X$  is the visible domain discriminator that can be used to identify the authenticity of visible image translated by infrared image and real visible image. For the generating network from visible image to infrared image  $G_{X \rightarrow Y}$ , the generative adversarial loss function is as follows:

$$\min_G \max_D L_{GAN}(G_{X \rightarrow Y}, D_Y, X, Y) = E_{y \sim p_{data}(y)} [\log D_Y(y)] + E_{x \sim p_{data}(x)} [\log(1 - D_Y(G_{X \rightarrow Y}(x)))] \quad (1)$$

where  $x \sim P_{data}(x)$  is from the real distribution data of visible domain,  $y \sim P_{data}(y)$  is from the real distribution data of infrared domain, and  $E[\cdot]$  is the mathematical expectation.

The cycle consistency loss is used to limit the reconstruction error of the two generating networks at pixel level, the loss function is as follows:

$$L_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, X, Y) = E_{x \sim p_{data}(x)} [\|G_{Y \rightarrow X}(G_{X \rightarrow Y}(x)) - x\|_1] + E_{y \sim p_{data}(y)} [\|G_{X \rightarrow Y}(G_{Y \rightarrow X}(y)) - y\|_1] \quad (2)$$

where  $\|\cdot\|_1$  is one-dimensional Euclidean distance.

In addition, the identity loss is used to keep the color consistency between the translated image and the original image.

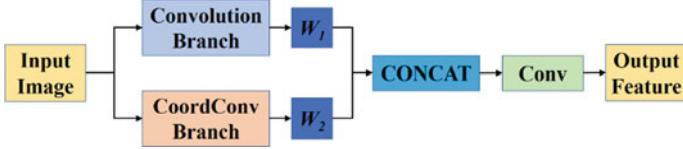
$$L_{ident}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, X, Y) = E_{x \sim p_{data}(x)} [\|G_{X \rightarrow Y}(y) - y\|_1] + E_{y \sim p_{data}(y)} [\|G_{Y \rightarrow X}(x) - x\|_1] \quad (3)$$

It can be seen from the above equations that CycleGAN only uses pixel-level alignment through one-dimensional Euclidean distance, and lacks the semantic supervision information of the generated image. In addition, it has a better generation effect for objects with simple geometry, such as ocean and sky, but not for objects with specific geometry.

### 3 Proposed Method

#### 3.1 Two-Way Feature Fusion

Two-way feature fusion structure uses different feature extraction methods for each Branch. One uses traditional convolution, and the other uses CoordConv [8]. The fusion operation uses concatenation and convolution to generate the output feature. By designing learnable weights  $W_1$  and  $W_2$ , the network can adaptively find the



**Fig. 2** Two-way feature fusion structure

optimal feature fusion method and improve the feature extraction ability of the generated network (Fig. 2).

### 3.2 Feature-Level Alignment

Deep learning network contains image detail information in shallow layer and image semantic information in deep layer. In order to make the generated image have more accurate semantic information, the images in the dataset are first used to pretrain the YOLO object detection network [9]. After the network achieves good results, the network parameters are frozen to assist the image generation process. The feature map after three times down-sampling of YOLO backbone network is extracted, and align the deep semantic features of the generated image with the original image. The perceptual loss function is as follows:

$$\begin{aligned}
 L_{perc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, F_{YOLO}, X, Y) \\
 = E_{x \sim p_{data}(x)} [\|F_{YOLO}(G_{Y \rightarrow X}(G_{X \rightarrow Y}(x))) - F_{YOLO}(x)\|_2] \\
 + E_{y \sim p_{data}(y)} [\|F_{YOLO}(G_{X \rightarrow Y}(G_{Y \rightarrow X}(y))) - F_{YOLO}(y)\|_2]
 \end{aligned} \quad (4)$$

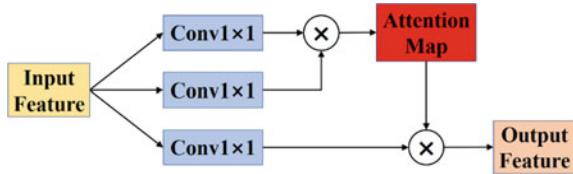
where  $F_{YOLO}$  is the deep feature map of YOLO backbone network, and  $\|\cdot\|_2$  is the two-dimensional Euclidean distance.

The circularly generated image should have the same semantic information as the image output from the regularization process, namely:

$$\begin{aligned}
 L_{sem}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, X, Y) = E_{x \sim p_{data}(x)} [\|F_{Y \rightarrow X}(x) - F_{Y \rightarrow X}(G_{X \rightarrow Y}(x))\|_2] \\
 + E_{y \sim p_{data}(y)} [\|F_{X \rightarrow Y}(y) - F_{X \rightarrow Y}(G_{Y \rightarrow X}(y))\|_2]
 \end{aligned} \quad (5)$$

where  $F_{X \rightarrow Y}$  and  $F_{Y \rightarrow X}$  are top-level feature for generating networks  $G_{X \rightarrow Y}$  and  $G_{Y \rightarrow X}$  respectively.

**Fig. 3** Self-attention module



### 3.3 Self-Attention Mechanism

Convolution operation is the main method of feature extraction in CycleGAN, while convolution operation only deals with local neighbourhood information in the image, and some global information in the image is missed, which affects the effect of image generation. The self-attention module consists of convolution operation and matrix multiplication. By introducing the self-attention mechanism in the generator, all features position in the image can be used to generate more detailed images (Fig. 3).

After adding self-attention mechanism, the output feature map is as follows:

$$F_{out} = x + \gamma \cdot SA(x) \quad (6)$$

where  $SA()$  is the self-attention module,  $\gamma$  is the learnable weight, which has an initial value of 0 at the beginning of training and gradually learns the weight of foreground and background information, and  $x$  is the input feature map.

### 3.4 Loss Function

The loss function of CycleGAN is as follows:

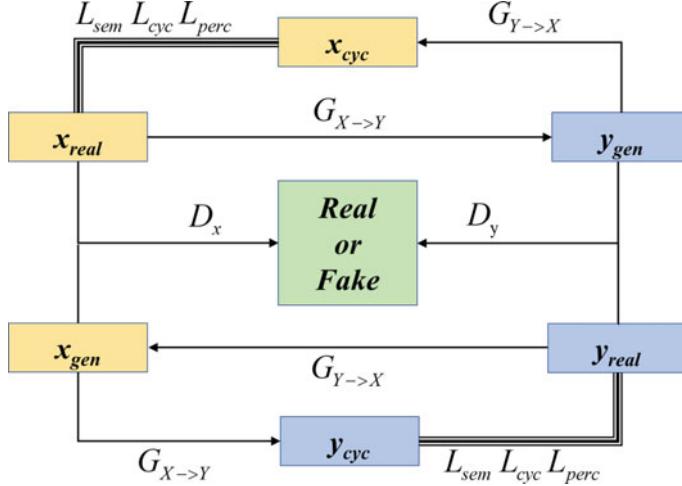
$$\begin{aligned} L_{CycleGAN}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, D_X, D_Y, X, Y) \\ = \lambda_{GAN} L_{GAN}(G_{X \rightarrow Y}, D_Y, X, Y) + \lambda_{GAN} L_{GAN}(G_{Y \rightarrow X}, D_X, X, Y) \\ + \lambda_{cyc} L_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, X, Y) + \lambda_{ident} L_{ident}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, X, Y) \end{aligned} \quad (7)$$

where  $\lambda_{GAN}$ ,  $\lambda_{cyc}$  and  $\lambda_{ident}$  are the loss weights of different parts respectively.

In combination with Eqs. (4, 5, 7), the new loss function can be expressed as (Fig. 4):

$$\begin{aligned} L_{total} = L_{CycGAN}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, D_X, D_Y, X, Y) \\ + \lambda_{perc} L_{perc}(G_{x \rightarrow y}, G_{y \rightarrow x}, N_{YOLO}, X, Y) \\ + \lambda_{sem} L_{sem}(G_{x \rightarrow y}, G_{y \rightarrow x}, X, Y) \end{aligned} \quad (8)$$

where  $\lambda_{perc}$  and  $\lambda_{sem}$  are the loss weights of different parts respectively.



**Fig. 4** The improved image-to-image translation network

## 4 Experimental Results

### 4.1 Experimental Setting

The experiment is carried out on a PC with Quadro RTX 8000 (48 GB storage) GPU and Intel Core i9-10900 k CPU. The operating system is Ubuntu 20.04 LTS, CUDA version is 11.0, cudnn version is 8.0, and the deep learning framework uses Pytorch 1.6. The optimizer uses Adam and the learning rate is  $10^{-4}$  until 250,000 iterations and linearly decayed up to 500,000 iterations. The training image size is  $512 \times 512$ . The weights of loss function are set to  $\lambda_{GAN} = 1$ ,  $\lambda_{cyc} = 10$ ,  $\lambda_{identity} = 10$ , which are consistent with CycleGAN. For more stable training, based on the convergence of the network, the weights  $\lambda_{perc} = 10$  and  $\lambda_{sem} = 0.05$  are set according to the corresponding part of the loss value and CycleGAN loss value in the same order.

### 4.2 Dataset and Evaluation Metric

In terms of the dataset, the low altitude sea surface visible and infrared dataset contains 4085 infrared images with the resolution of  $640 \times 512$ , and 4352 visible images with the resolution of  $640 \times 360$ , including ships and UAVs. In terms of evaluation metric, the lower the FID (Fréchet Inception Distance) score, the closer the generated image is to the real image. The equation of FID is as follows:

$$FID = \|\mu_r - \mu_g\|^2 + Tr \left[ \Sigma_r + \Sigma_g - 2 \times (\Sigma_r \Sigma_g)^{\frac{1}{2}} \right] \quad (9)$$

**Table 1** Improvement effect of image-to-image translation network

Baseline (CycleGAN)	Two-way feature fusion	Feature-level alignment	Self-attention mechanism	FID
✓				52.71
✓	✓			47.30
✓	✓	✓		40.45
✓	✓	✓	✓	36.93

where  $\mu_r$  and  $\mu_g$  are the mean value of the feature vector of the real image and generated image respectively,  $\Sigma_r$  and  $\Sigma_g$  are the feature covariance matrix of the real image and generated image respectively, and  $Tr[\cdot]$  is the trace of the matrix.

### 4.3 Ablation Study

**Improvement effect of image-to-image translation network.** It can be seen from Table 1 that the FID score of the CycleGAN is higher, because only the pixel-level loss function is used to constrain the image generation. The semantic consistency of the generated image is ensured by using perceptual loss and semantic loss, and the self-attention mechanism makes the generation network pay attention to the global feature information, so as to distinguish the foreground–background information more effectively and improve the image generation (Fig. 5). Compared with baseline, the improved network FID score decreased by 15.78.

**Data augmentation.** Through the improved image-to-image translation network, the visible image in the dataset is translated into the infrared image, so as to make infrared image data augmentation and further improve the infrared object detection accuracy. The results before and after infrared image data augmentation are shown in Table 2. The number of images increased by 2709, the mAP@0.5 increases by 5.28%.

## 5 Conclusion

In this work, an improved image-to-image translation network is designed and used in infrared image data augmentation. Two-way feature fusion is designed to enhance feature extraction ability to generate network, and the feature consistency of generated image is ensured by perceptual loss and semantic loss. The self-attention mechanism enables the generation network to pay attention to the global information of the image. The results show that the proposed method can effectively improve the quality of image generation, and the FID score is reduced by 15.78. In the aspect of



**Fig. 5** Display of infrared image translated from visible image (The second line of infrared image is generated by the proposed method, and the third line of infrared image is generated by CycleGAN)

**Table 2** Results of the infrared image data augmentation

Data augmentation	Number of training images	mAP@0.5 (%)
Before	3000	88.31
After	5709	93.59

infrared data augmentation, combined with the generated infrared image data, the number of training images increased by 2709, the mAP@0.5 increases by 5.28%, which effectively enhances the image diversity of the training dataset and improves the object detection accuracy.

## References

1. Ian, J.G., Jean, P.A., Mehdi, M., Bing, X., David, W.F., Sherjil, O., Aaron, C., Yoshua, B.: Generative adversarial networks. arXiv preprint [arXiv:1406.2661](https://arxiv.org/abs/1406.2661) (2014)
2. Changjin, L., Jian, C., Xing, Z.: Design and implementation of an infrared image generative model. In: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, pp. 1338–1345. IEEE, Dalian (2020)

3. Zhixiang, W., Zheng, W., Yinqiang, Z., Yung-Yu, C., Shin'ichi, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 618–626. IEEE, Long Beach (2019)
4. Guan'an, W., Tianzhu, Z., Jian, C., Si, L., Yang, Y., Zengguang, H.: RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623–3632. IEEE, Seoul (2019)
5. Foji, C., Feng, Z., Qingxiao, W., Yingming, H., Ende, W.: Infrared image data augmentation based on generative adversarial network. *J. Comput. Appl.* **40**(7), 2084–2088 (2020)
6. Jun-Yan, Z., Taesung, P., Phillip, I., Alexei A.E.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232. IEEE, Venice (2017)
7. Linmiao, H., Yong, Z.: Facial image translation in short-wavelength infrared and visible light based on generative adversarial network. *Acta Optica Sinica* **40**(05), 75–84 (2020)
8. Rosanne, L., Joel, L., Piero, M., Felipe Petroski, S., Eric, F., Alex, S., Jason, Y.: An intriguing failing of convolutional neural networks and the CoordConv solution. arXiv preprint [arXiv:1807.03247](https://arxiv.org/abs/1807.03247) (2018)
9. Alexey B., Chien-Yao, W., Hong-Yuan Mark, L.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)

# Rotation Invariant Convolutional Neural Network Based on Orientation Pooling and Covariance Pooling



Xiaoqin Yao , Tiecheng Song , Jingying Zeng , and Yangming Xie

**Abstract** Convolutional neural networks (CNNs) have shown promising performance in a variety of visual tasks. However, by using convolution operations, CNNs have limited ability to handle orientation changes of input images. To alleviate this problem, data augmentation is typically used. However, this solution brings additional computational and storage costs and has no sufficient theoretical guarantees on rotation invariance. Alternative solutions have been proposed which only take the first-order features into consideration, without utilizing the higher-order information for representation learning. In this paper, we propose an end-to-end rotation invariant CNN (RICNN) based on orientation pooling and covariance pooling to classify rotated images. Specifically, we learn deep rotated filters to extract rotation invariant feature maps by using two types of orientation pooling (OP), including max OP and average OP. Furthermore, we employ covariance pooling to extract rotation invariant hierarchical second-order features. Experiments on two datasets demonstrate the effectiveness of RICNN for rotation invariant image classification.

**Keywords** Rotation Invariance · CNN · Pooling · Covariance · Classification

## 1 Introduction

The exploitation of invariant information using deep neural networks has become an important topic in the field of computer vision. For many visual recognition tasks, it is desirable to extract rotation invariant features from input data, e.g., biomedical microscopy and astronomical images, captured under different orientations.

Traditional methods focus on extracting handcrafted rotation invariant features such as Gabor filters [1], SIFT [2], RI-HOG [3] and RI-LBP [4]. The design of these handcrafted features requires careful engineering and expert knowledge, thereby limiting their wide application. It is well known that convolutional neural networks (CNNs) can learn expressive features directly from raw data. However, by using

---

X. Yao ( ) · T. Song · J. Zeng · Y. Xie

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

e-mail: [s190131089@stu.cqupt.edu.cn](mailto:s190131089@stu.cqupt.edu.cn)

convolution operations, CNNs have limited ability to process orientation changes of input images. To handle arbitrary visual transformations, data augmentation [5] is often employed. This solution can enhance the network performance by extending the dataset, but requires additional computational and storage costs. Also, this solution has no sufficient theoretical guarantees on transformation invariance.

To alleviate the above problems, Laptev et al. [6] use parallel network architectures and the TI-POOLING operator to obtain transformation invariant features. Jaderberg et al. [7] build the spatial transformer network (STN) to learn a canonical pose and generate an invariant representation through warping. Moreover, Esteves et al. [8] introduce the polar transformer network (PTN) to provide equivariance for scaling and rotation. Recent studies [9–15] have implemented the prior knowledge of rotation on the most fundamental elements of deep CNNs, i.e., convolution operators. Worrall et al. [11] design harmonic networks (H-Nets), a CNN that shows equivariance for translations and rotations by replacing traditional CNN filters with circular harmonics. Romero et al. [13] propose attentive group equivariant convolutions where attention is employed to emphasize meaningful symmetric combinations and suppress unreasonable and misleading symmetric combinations during the course of convolution. In addition, Zhou et al. [14] design actively rotating filters (ARFs) to produce feature maps by encoding orientation and position information. Marcos et al. [15] present rotation equivariant vector field network (RotEqNet), which extracts a vector field from the feature maps to achieve rotational equivariant structure.

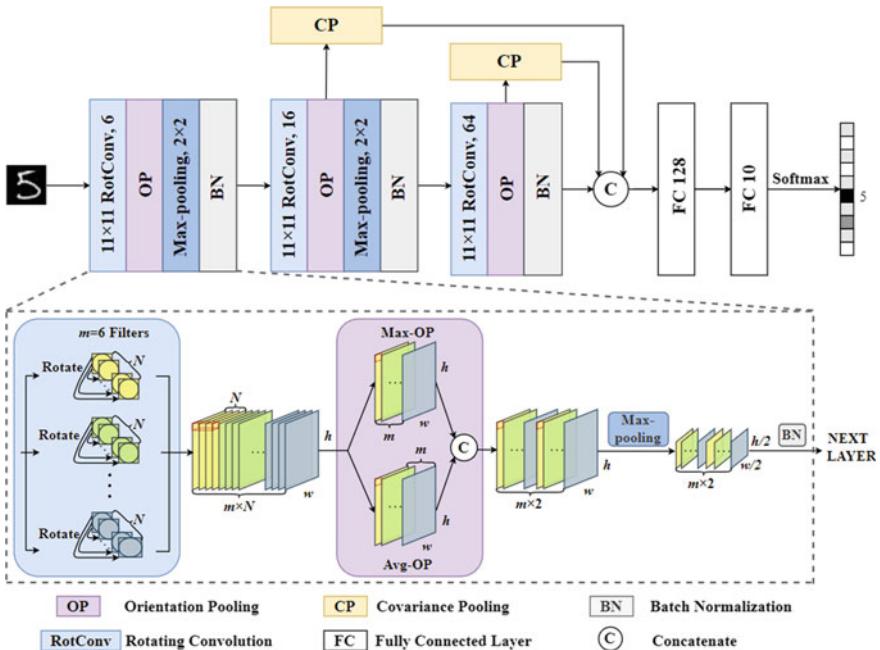
The above methods only take the first-order features into consideration and do not make use of the higher-order information for representation learning [16–18]. High-order features have been shown to be effective for feature description. For example, Dai et al. [19] propose to combine the first-order information derived by averaging pooling and the second-order information derived by a bilinear model [20] through a concatenation operation. In [21, 22], He et al. propose to combine feature maps from different convolutional layers and use covariance pooling to achieve more discriminative feature learning. However, the extraction of deep higher-order representations that are invariant to image rotation is not addressed in these methods.

In view of the above problems, we apply rotating convolution operator to obtain multi-orientation filter responses and employ two types of orientation pooling (OP), i.e., max OP and average OP, to achieve rotation invariance. Furthermore, we embed covariance pooling to learn second-order rotation invariant features. In this way, we construct a rotation invariant CNN model which shows good performance for image classification. We summarize the contributions of this paper as follows:

- We develop a CNN architecture capable of learning a rotation invariant image representation for image classification.
- We propose to use orientation pooling (including max OP and average OP) and covariance pooling for representation learning. These pooling strategies not only fully exploit first- and second-order information but also have invariance to image rotation.

## 2 Method

Figure 1 shows the framework of RICNN for image classification. First, the  $m$  canonical filters are rotated, each forming  $N$  rotated versions at the directions from 0 to 360 degrees. Then, we convolve the input image with the rotated filters to obtain a total of  $m \times N$  feature maps. Next, we pass the feature maps to the orientation pooling (OP) module (including max OP and average OP) to learn rotation invariant features, followed by max-pooling and batch normalization (BN). The above process is repeated to learn deep invariant features. Meanwhile, we feed the output features of the OP module into the covariance pooling (CP) module and concatenate the results as the input of the fully connected layer (FC). In this way, the learned features contain first- and second-order rotation invariant features which are discriminative for image classification. Now, we elaborate the process of the rotating convolution, the orientation pooling, and the covariance pooling.



**Fig. 1** Framework of the proposed RICNN for image classification ( $h$  and  $w$  are the size of the feature map.  $m$  is the number of canonical filters.  $N$  is the number of orientations)

## 2.1 Rotating Convolution (RotConv)

We apply rotation to the learnable convolutional filters to obtain multi-orientation feature maps [14, 15]. Specifically, given a canonical filter  $w \in \mathbb{R}^{s \times s \times d}$  where  $s \times s$  is the size of canonical filter and  $d$  is the number of channels, we rotate this filter to  $N$  orientations in the spatial domain. The  $k$ -th rotated version of the canonical filter  $w$  is computed as

$$w^k = f_{\theta_k}(w) \quad (1)$$

$$\theta_k = k \frac{360}{N}, \quad \forall k = 0, 1, \dots, N - 1 \quad (2)$$

where  $f_{\theta_k}$  is the  $\theta_k$ -degree rotation operator,  $N$  is the total number of rotation angles and  $w^k$  is computed by bilinear interpolation of the resampling of  $w$  after rotating  $\theta_k$  degrees around the center of the convolution kernel. Since rotation may cause weights near the corners of the convolution kernel to be out of the effective range, we only use the weights within a circle with a diameter of  $s$  pixels to compute the convolution.

Given an input image  $x$ , we apply  $p$  zero-padding to it, i.e.,  $x \in \mathbb{R}^{(h+p) \times (w+p) \times d}$ . In this paper, we set  $p = \text{floor}(s/2)$  and  $s = 11$ . For each canonical filter, we convolve the image with the rotated filters to obtain  $N$  multi-orientation feature maps  $y \in \mathbb{R}^{h \times w \times N}$  computed as

$$y^{(k)} = x * w^k, \quad \forall k = 0, 1, \dots, N - 1 \quad (3)$$

where  $*$  is the convolution operator.

During the back-propagation, the gradients with respect to each rotated convolution filter  $\nabla w^k$  are aligned to the canonical form and summed as follows

$$\nabla w = \sum_k f_{-\theta_k}(\nabla w^k) \quad (4)$$

where  $f_{-\theta_k}$  is the alignment operator which is an inverse of  $f_{\theta_k}$ .

## 2.2 Orientation Pooling (OP)

For a canonical filter, its output tensor  $y \in \mathbb{R}^{h \times w \times N}$  passes through orientation pooling, returning the maximum and average values of the activations for all rotation orientation at each spatial location. The resulting feature maps for all canonical filters are concatenated and fed into the max pooling operation (i.e., subsampling).

Specifically, given the output tensor  $y \in \mathbb{R}^{h \times w \times N}$ , the maximum activation magnitudes  $\chi_1 \in \mathbb{R}^{h \times w \times d_1}$  and average activation magnitudes  $\chi_2 \in \mathbb{R}^{h \times w \times d_2}$  for pixel  $[i, j]$  are derived as follows

$$\chi_1[i, j] = \max_{k=0 \dots N-1} y[i, j, k] \quad (5)$$

$$\chi_2[i, j] = \frac{1}{N} \sum_{k=0}^{N-1} y[i, j, k] \quad (6)$$

Note that only one canonical filter is illustrated here and hence  $d_1 = d_2 = 1$ . Subsequently, the aggregated feature map  $\chi$  is obtained by concatenating  $\chi_1$  and  $\chi_2$  as follows

$$\chi = [\chi_1; \chi_2] \in \mathbb{R}^{h \times w \times d} \quad (7)$$

where  $[, ; ,]$  denotes the concatenation along the channel dimension and  $d = d_1 + d_2$ .

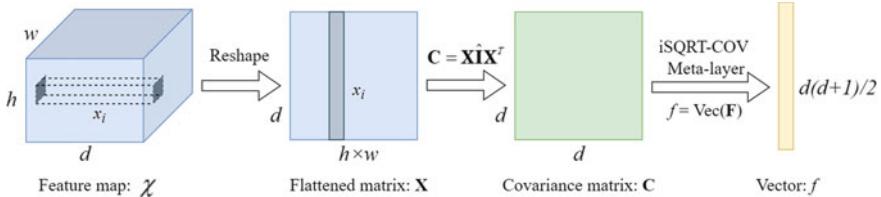
From Eqs. (5) and (6), one can see that the proposed OP uses both the average and maximal values which can capture distinct statistical features while achieving the rotation invariance.

### 2.3 Covariance Pooling (CP)

In this module, we exploit covariance pooling to capture global second-order features. This allows the CNN model to learn the correlation information from different filter responses and obtain complementary features to enhance classification performance.

Let the feature map  $\chi \in \mathbb{R}^{h \times w \times d}$  be the aggregated results of OP for  $m$  canonical filters. First, we reshape  $\chi$  into a matrix  $X \in \mathbb{R}^{d \times n}$  with  $n = h \times w$ , as shown in Fig. 2. Then, we compute the covariance matrix  $\mathbf{C}$  as follows

$$\mathbf{C} = X \hat{X}^T \in \mathbb{R}^{d \times d} \quad (8)$$



**Fig. 2** Illustration of the proposed covariance pooling module

where  $\hat{\mathbf{I}} = 1/n(\mathbf{I} - 1/n\mathbf{1})$ ,  $\mathbf{I}$  is an  $n \times n$  identity matrix, and  $\mathbf{1}$  is a matrix with all elements being 1. Since all the elements in each  $x_i$  are derived from the rotation invariant feature maps, the pooling covariance matrix in Eq. (8) is invariant to image rotation.

Then, we introduce the iSQRT-COV meta-layer [18], which is used to approximate the square root of the covariance matrix. The eigen-decomposition (EIG) of covariance matrix  $\mathbf{C}$  is calculated as

$$\mathbf{C} = \mathbf{U}\text{diag}\{\lambda_i\}\mathbf{U}^T, \quad i = 1, \dots, d \quad (9)$$

where  $\text{diag}\{\lambda_i\}$  and  $\mathbf{U}$  are the eigenvalue matrix and eigenvector matrix of  $\mathbf{C}$ , respectively. The square root of matrix  $\mathbf{C}$  is calculated as

$$\mathbf{F} = \mathbf{U}\text{diag}\left\{\lambda_i^{1/2}\right\}\mathbf{U}^T, \quad i = 1, \dots, d \quad (10)$$

Equation (10) regularizes the square root of the covariance matrix. Since EIG is not well supported by GPU, the iSQRT-COV meta-layer uses an iterative method to approximate the square root of the covariance matrix. Note that the output  $\mathbf{F}$  is a symmetric matrix and we only need to concatenate its upper triangular entries to obtain a vector  $f$  of dimension  $d(d + 1)/2$ .

### 3 Experiments

#### 3.1 Experimental Settings and Datasets

For the proposed RICNN, we perform training using Adam [23] as the optimizer, starting with a learning rate of 0.1 and gradually decreasing to 0.001. The training epochs and batch size are 45 and 32, respectively. The classification objective is the commonly used cross-entropy function. The number of orientations is set to  $N = 17$ . We run the experiments with GeForce GTX TITAN X (12G). We evaluate the performance of RICNN on two datasets: MNIST-rot-12k and FMNIST-rot-12k.

**MNIST-rot-12k.** The dataset is composed of  $28 \times 28$ ,  $360^\circ$  rotated images of handwritten digits (0–9). It consists of 10k training samples, 2k validation samples and 50k test samples.

**FMNIST-rot-12k.** The dataset contains 10 classes of grayscale clothing images ( $28 \times 28$  pixels). Similarly, it has 10k training samples, 2k validation samples and 50k test samples, where the images of the original FMNIST are rotated by angles between  $[0^\circ, 360^\circ]$ .

**Table 1** Results on the MNIST-rot-12k and FMNIST-rot-12k

Method	Error (%)	
	MNIST-rot-12k	FMNIST-rot-12
SVM	10.38	–
TIRBM	4.20	24.46
H-Net	2.14	14.50
DREN	1.56	–
ORN-8 (ORAlign)	1.63	13.42
$\alpha$ -p4-CNN	1.68	14.65
RotEqNet	1.31	13.39
TI-pooling	1.37	12.54
Ours (without CP)	1.20	13.18
Ours	<b>1.07</b>	<b>12.24</b>

### 3.2 Experimental Results

Table 1 shows the classification results of different methods including SVM [24], TIRBM [10], H-Net [11], DREN [25], ORN [14],  $\alpha$ -p4-CNN [13], RotEqNet [15] and TI-pooling [6]. The results of SVM, TIRBM and DREN are quoted from the original papers and other results are reported based on our own implementation using the codes provided by the authors.

From Table 1 we can make the following observations. Firstly, our RICNN achieves the best performance on both datasets, demonstrating the robustness of our network model for image rotation. Secondly, TI-pooling and RotEqNet which consider only first-order information show worse performance than the proposed RICNN. By contrast, our RICNN can obtain better performance by additionally using covariance pooling to extract hierarchical second-order statistical features. Finally, we report the results of RICNN with and without covariance pooling. It can be observed that RICNN without covariance pooling leads to a decrease in accuracy. Hence, the proposed second-order covariance pooling complements the first-order orientation pooling for rotation invariant feature learning.

### 3.3 Ablation Study

**Impacts of convolution kernel size.** Table 2 shows the classification results of RICNN by varying the convolution kernel size. As can be seen, better results are obtained on MNIST-rot-12k and FMNIST-rot-12k with the kernel size  $11 \times 11$ . When the convolution kernel size is too small or too large, the performance can drop slightly. This is because the small convolutional kernels hardly capture large scale image structures while the large ones can introduce noisy image patterns. By

**Table 2** Impacts of convolution kernel size on the classification results

kernel size	Error (%)	
	MNIST-rot-12k	FMNIST-rot-12k
$7 \times 7$	1.26	13.17
$9 \times 9$	1.15	12.39
$11 \times 11$	<b>1.07</b>	<b>12.24</b>
$13 \times 13$	1.21	12.75

**Table 3** Impacts of the number of orientations on the classification results

$N$	Error (%)	
	MNIST-rot-12k	FMNIST-rot-12k
9	1.19	13.08
13	1.16	12.65
17	<b>1.07</b>	<b>12.24</b>
21	1.17	12.45

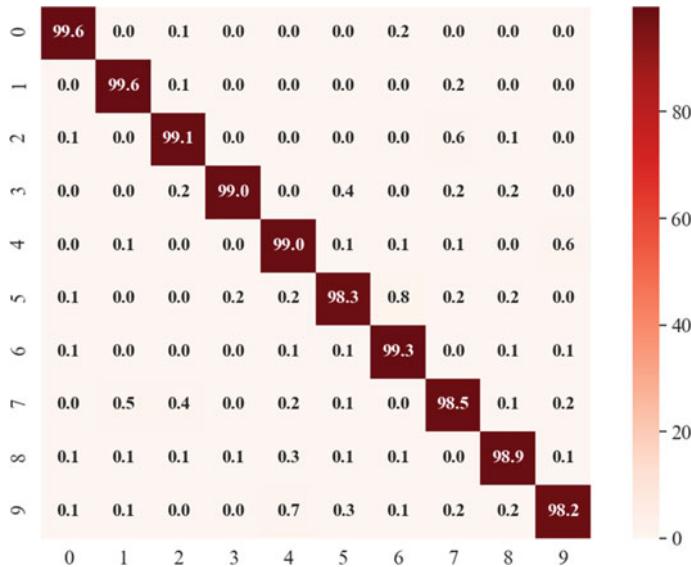
considering the effectiveness and efficiency, in this work we have chosen the kernel size  $11 \times 11$  for both datasets.

**Impacts of the number of orientations.** Table 3 shows the classification performance of RICNN by varying the number of orientations. As can be seen from Table 3, the best classification results are obtained by using  $N = 17$  on both datasets. There is no improvement in classification results as the number of orientations increases to 21. This is because the network has acquired enough orientation information when  $N = 17$ . To ensure both the effectiveness and efficiency, we have chosen  $N = 17$  in our experiments.

**Comparison of orientation pooling strategies.** Table 4 compares the classification results of RICNN using different orientation pooling strategies. Here, “maximum” and “average” denote max-pooling and average-pooling applied to the orientation dimensions, returning the maximum value and average value at each spatial location, respectively. “Addition” represents adding the obtained maximum value and average value at each spatial location. “Concatenation” denotes concatenating the maximum value and average value of the feature maps. As can be seen from Table 4, our

**Table 4** Classification results with different pooling strategies

Pooling	Error (%)	
	MNIST-rot-12k	FMNIST-rot-12k
Average	8.42	22.51
Maximum	1.18	12.56
Addition	1.22	12.43
Concatenation	<b>1.07</b>	<b>12.24</b>



**Fig. 3** Confusion matrix of RICNN on the MNIST-rot-12k

concatenation operator obtains the best classification results on both datasets, which demonstrates the efficacy of the proposed orientation pooling operator.

**Confusion matrix.** Figure 3 further shows the confusion matrix for the 10 categories on the MNIST-rot-12 k. As shown in Fig. 3, the digits “0” and “1” have the highest classification accuracy of 99.6%. The two most confused cases are that “5” is taken for “6” (0.8%) and “9” is taken for “4” (0.7%). Misclassification is due to the structural similarity between these image pairs.

## 4 Conclusions

We propose RICNN, a rotation invariant end-to-end CNN model, based on orientation pooling and covariance pooling for image classification. Instead of only taking the first-order information into consideration, we employ covariance pooling (CP) to generate second-order statistical features. We also adopt two types of orientation pooling (OP), including max OP and average OP, to aggregate richer features. The resulting feature representation is rotation invariant and discriminative. Experiments on two benchmark datasets demonstrate the effectiveness of RICNN for rotation invariant image classification.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (61702065), Chongqing Research Program of Basic Research and Frontier Technology (cstc2018jcyjAX0033), and Natural Science Foundation of Chongqing (cstc2020jcyjmsxmX0636).

## References

1. Haley, G. M., Manjunath, B. S.: Rotation-invariant texture classification using modified Gabor filters. In: Proceedings of International Conference on Image Processing, vol. 1, pp. 262–265. IEEE, Washington (1995)
2. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE, Greece (1999)
3. Liu, K., Skibbe, H., Schmidt, T., Blein, T., Palme, K., Brox, T., Ronneberger, O.: Rotation-invariant HOG descriptors using Fourier analysis in polar and spherical coordinates. *Int. J. Comput. Vision* **106**(3), 342–364 (2014)
4. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
5. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: learning augmentation strategies from data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 113–123. IEEE, Long Beach (2019)
6. Laptev, D., Savinov, N., Buhmann, J. M., Pollefeys, M.: TI-POOLING: transformation-invariant pooling for feature learning in convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 289–297. IEEE, Las Vegas, NV (2016)
7. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025. Montreal (2015)
8. Esteves, C., Allen-Blanchette, C., Zhou, X., Daniilidis, K.: Polar transformer networks. arXiv preprint [arXiv:1709.01889](https://arxiv.org/abs/1709.01889) (2017)
9. Kivinen, J. J., Williams, C.K.: Transformation equivariant Boltzmann machines. In: International Conference on Artificial Neural Networks, pp. 1–9. Springer, Heidelberg (2011)
10. Sohn, K., Lee, H.: Learning invariant representations with local transformations. In: 29th International Conference on Machine Learning, pp. 1311–1318. Scotland (2012)
11. Worrall, D.E., Garbin, S. J., Turmukhambetov, D., Brostow, G. J.: Harmonic networks: Deep translation and rotation equivariance. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5028–5037. IEEE, Honolulu, HI (2017)
12. Weiler, M., Hamprecht, F. A., Storath, M.: Learning steerable filters for rotation equivariant CNNs. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 849–858. IEEE, Salt Lake City, UT (2018)
13. Romero, D., Bekkers, E., Tomczak, J., Hoogendoorn, M.: Attentive group equivariant convolutional networks. In: 37th International Conference on Machine Learning, pp. 8188–8199 (2020)
14. Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Oriented response networks. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 519–528. IEEE, Honolulu, HI (2017)
15. Marcos, D., Volpi, M., Komodakis, N., Tuia, D.: Rotation equivariant vector field networks. In: 16th IEEE International Conference on Computer Vision (ICCV), pp. 5048–5057. IEEE, Venice (2017)
16. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Free-form region description with second-order pooling. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(6), 1177–1189 (2014)
17. Acharya, D., Huang, Z., Pani Paudel, D., Van Gool, L.: Covariance pooling for facial expression recognition. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 367–374. IEEE, Salt Lake City (2018)
18. Li, P., Xie, J., Wang, Q., Gao, Z.: Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 947–955. IEEE, Salt Lake City (2018).
19. Dai, X., Yue-Hei Ng, J., Davis, L.S.: Fason: first and second order information fusion network for texture recognition. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7352–7360. IEEE, Honolulu, HI (2017)

20. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1449–1457. IEEE, Santiago (2015)
21. He, N., Fang, L., Li, S., Plaza, A., Plaza, J.: Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Trans. Geosci. Remote Sens.* **56**(12), 6899–6910 (2018)
22. He, N., Fang, L., Li, S., Plaza, J., Plaza, A.: Skip-connected covariance network for remote sensing scene classification. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(5), 1461–1474 (2019)
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
24. Larochelle, H., Erhan, D., Courville, A., Bergstra, J., Bengio, Y.: An empirical evaluation of deep architectures on problems with many factors of variation. In: 24th International Conference on Machine Learning, pp. 473–480. Corvallis (2007)
25. Li, J., Yang, Z., Liu, H., Cai, D.: Deep rotation equivariant network. *Neurocomputing* **290**, 26–33 (2018)

# Deep Spatial–Temporal Graph Modeling of Urban Traffic Accident Prediction



Yongxian Huang , Fan Zhang , and Jinhui Hu

**Abstract** In recent years, more and more attention is attracted to analyzing the factors that affect and induce traffic accidents, and studying the methods of traffic accident prediction, which may help reduce the risk of accidents and alleviate traffic congestion. In many works, urban space is divided into small pieces and represented as a graph structure composed of nodes (vertices) and edges, and then graph convolutional networks are used to analyze the spatial correlation characteristics of traffic flow parameters. In this paper, a computer-vision based method is adopted to calculate the similarity of nodes in urban road network image as an auxiliary term of adjacency matrix. Moreover, a gated graph convolutional multi task (GGCMT) model is proposed to predict the accident risks on both node scale and city scale (overall risk). In order to reduce the interference of data sparsity, we propose an improved prior risk data enhancement method. The results show that the proposed model outperforms the baseline models. In the auxiliary task, the predicted overall accident risks are also consistent with the real values.

**Keywords** Traffic accident prediction · Urban traffic · Graph convolutional network

## 1 Introduction

Urban road traffic is a complex nonlinear system. It is difficult to accurately describe some of the rules through mathematical equations, such as the evolution of traffic flow in the road network, the probability distribution of traffic accident risk, and so on. With the help of traffic big data and deep learning, real-time traffic flow prediction [1], arrival time estimation [2], OD prediction [3], etc., have been realized. At the

---

Y. Huang · F. Zhang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, People's Republic of China  
e-mail: [yongxian@mail.ustc.edu.cn](mailto:yongxian@mail.ustc.edu.cn)

Y. Huang · J. Hu

The Smart City Research Institute of China Electronics Technology Group Corporation, Shenzhen, Guangdong 518033, People's Republic of China

same time, the development of traffic big data and deep learning can also provide a reliable solution for traffic accident risk prediction.

Many researches try to use deep learning model to solve the problem of traffic accident analysis. Some of them analyze the data in Euclidean domains [4–7], using convolutional neural networks (CNNs) to analyze urban traffic characteristics. However, the urban traffic network data have the characteristics of non-Euclidean structure, which may not be well described by convolutional neural networks.

A common method to analyze the non-Euclidean structure data in traffic prediction or other hot events prediction is graph convolutional neural networks [8–10]. In many works [11, 12], urban space is divided into regular grid areas as nodes of graph structure, and the connection between grids is regarded as edge. However, to our knowledge, there is no universally accepted method to calculate adjacency matrix. To fully consider the characteristics of traffic network, such as the directions of traffic roads and distribution density of road network, in addition to calculating the static characteristics of the road network, we adopt a computer-vision method to calculate the visual similarity between nodes as an auxiliary adjacency matrix. The calculated adjacency matrix can not only manifest the internal characteristics of grids, but also the conditions around them.

One other obstacle may be in front when dealing with the sparse traffic accident data. When there exist too many zeros in the training labels, the zero-inflated problem will reduce the prediction accuracy of the model [13]. To tackle this issue, there are mainly two methods, respectively handling the loss function [9] and data preprocessing [11]. In this paper, an improved prior risk data enhancement method is applied in data preprocessing, considering not only the accident risk distribution probability of each node in history, but also the time-varying global accident risks.

In this paper, we study the problem of city-wide traffic accident prediction by proposing a gated graph convolutional multi task (GGCMT) framework. In addition to predicting the traffic accident risk of each node, we also predict the overall risk of the city in the auxiliary task. Compared with the previous model frameworks, our model shows better performance in accident risk prediction.

The paper is organized as follows. In Sect. 2, we introduce the multi-task and some basic definitions. Section 3 presents data preprocessing and the proposed framework GGCMT. Empirical studies are presented and discussed in Sect. 4. Finally, the conclusions of this paper will be in Sect. 5.

## 2 Multi-task and Definition

In this section, some basic definitions and the multi-task are presented. We first divide the study area (NYC without Staten Island) into small regular squares of size 1.5 km and obtain  $N = 383$  square regions  $v_1, v_2, \dots, v_N$ . Areas without road network distribution (occupied by mountains or water areas) are not taken into account, as there is no traffic activity.

## 2.1 Urban Graph

The urban area can be expressed as an undirected graph  $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ , where  $\mathbf{V}$ ,  $\mathbf{E}$ , and  $\mathbf{W}$  indicate the nodes, edges and weights of the graph, respectively. The square urban regions are represented as a set of graph nodes  $\mathbf{V}$ , and the connection between any pair of regions  $v_i$  and  $v_j$  is turned into edge  $e_{ij}$ . The weight  $w_{ij}$  corresponding to graph edge  $e_{ij}$  indicates the similarity or strength between the two vertices it connects.

## 2.2 Traffic Accident Risk

The value of traffic accident risk is defined as the sum of severity levels of accidents. In particular,  $r_{vi}(\Delta t) = \sum_{j=1}^3 j * \tau_{vi}^{\Delta t}(j)$ , where  $j$  indicates the severity level of accident, and  $\tau_{vi}^{\Delta t}(j)$  denotes the number of accidents of level  $j$ . The accident risk within time interval  $\Delta t$  can be represented by  $\mathbf{R}(\Delta t) = \{r_{v1}(\Delta t), r_{v2}(\Delta t), \dots, r_{vN}(\Delta t)\}$ . To analyze the development tendency of entire urban domain risk, we calculate the overall risk as  $\mathbf{R}_o(\Delta t) = \sum_i r_{vi}(\Delta t)$ .

Three accident risk levels are defined: minor accidents, injured accidents, and fatal accidents [14], whose weights are 1, 2, and 3, respectively.

## 2.3 Multi-task Traffic Accident Prediction

Given the urban road network features  $\mathbf{S} = \{S_{v1}, S_{v2}, \dots, S_{vN}\}$  and the history traffic accident risk  $\mathbf{R}(\Delta t)(\Delta t = 1, 2, \dots, T)$ , we can predict the distribution of the traffic accident risks  $\mathbf{R}(\Delta t)(\Delta t = T + 1, T + 2, \dots, T + L)$  and entire urban domain risk  $\mathbf{R}_o(\Delta t)(\Delta t = T + 1, T + 2, \dots, T + L)$ , where  $L$  is the number of intervals predicted.

# 3 Spatial-Temporal Traffic Accident Prediction

In this section, data preprocessing and the proposed framework GGCMT will be shown and elaborated.

### 3.1 Data Preprocessing

To obtain different weight matrices of the graph, we calculate the Jensen-Shannon divergence of the static road network characteristics and the computer-vision based similarity from urban road network image, respectively. In order to reduce the computational complexity, for each node, we select the most relevant nodes (nodes with the largest weight) in the urban graph as its adjacent node, keeping the connectivity ratio of the graph as  $\rho = 0.1$ .

To handle the sparse distribution characteristics of traffic accidents, we adopt an improved prior risk data enhancement method.

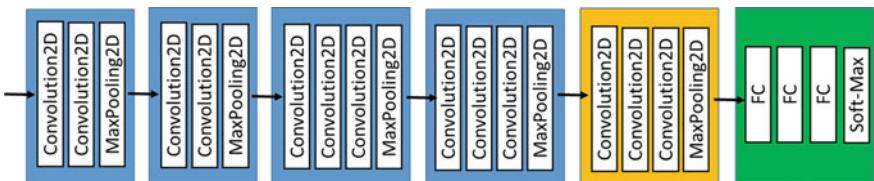
**Weight Matrix Derived from Static Road Network.** Corresponding to the division of urban areas, the static road network of NYC is divided into N parts. To calculate the similarity of static road characteristics, we express the characteristics of the static road network in each single node region as a vector. Then, the static road network weights matrix  $W_n$  can be calculated as [11]

$$w_{ij} = \begin{cases} 1, & \text{if region } v_i \text{ and } v_j \text{ are adjacent} \\ \exp(-JS(s_i|s_j)), & \text{otherwise} \end{cases} \quad (1)$$

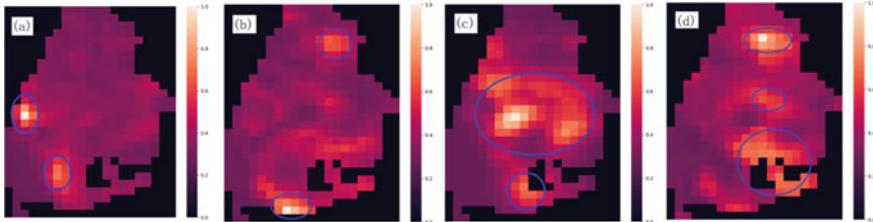
where  $s_i$  denotes the vector of static road network characteristics of node  $i$ .

**Weight Matrix Derived from Urban Road Network Image.** As the urban road image can visually demonstrate the road distribution density and other distribution characteristics in different areas of the city, we analyze its features by VGG16 model [15] pre-trained in ImageNet dataset. The model structure of VGG16 is illustrated in Fig. 1.

In image processing, the feature map in the front of convolutional neural network usually extracts the representative high-frequency and low-frequency features in the image; when the signal enters the deeper hidden layer, its more general and complete features will be extracted. Here, we extract the features of VGG16 model after the fourth pooling (the down sampling rate is 16) layer as the features of the node region, and calculate the cosine similarity between them as the weight of the corresponding edge.



**Fig. 1** Model structure of VGG16, layers in blue are used to calculate urban road network image features



**Fig. 2** Cosine similarity calculated with the visual features by VGG16. Different colors indicate the similarity between the features of each node and the target node (colored in white)

The calculated cosine similarity is illustrated in Fig. 2. We can observe that the distribution of similarity values in different patterns obviously distinguish from each other since their compared target nodes are different. For each target node, the similarity of nodes around it tends to be larger (in lighter color) than others in the distance. This is because that the receptive field [16] of the feature image exceeds the number of pixels occupied by the node region in the image when calculating the convolutional features, and it reflects the adjacency relationship between nodes to a certain extent. In Fig. 2c, the light color area near the target node is larger than others in Fig. 2a, b, d, illustrating there are more nodes with high similarity to the target node. Aside from the vicinity of the target node, some distant nodes will also appear lighter colors, indicating that there is a high similarity in the road network structure or density in these areas, and there may be some long-distance connection between them, such as similar urban functions and terrain.

Thus far, the weight matrix  $W_m$  reflects the similarity nodes in urban road network image is obtained.

**A Prior Risk Data Enhancement Method.** Since the observed accidents are usually few, the regression task tends to be the most frequent but uninteresting situation (no accidents). To solve this problem, Zhou et al. [11] proposed a prior risk data enhancement method, replacing zeros with the historical distribution of accident risk values of each node. However, it is time invariant, not able to reflect the time varying of the global accident risk. To this end, we proposed an improved prior risk data enhancement method.

We replace the zero values in risk data with a statistical accident risk index, whose value not only reflects the accident risk distribution probability of each region node in history, but also the time varying global accident risk distribution.

The statistical accident risk index is calculated as

$$I(\Delta t, v_i) = a(\Delta t) * \log p_i + b(\Delta t) \quad (2)$$

where  $p_i$  is the statistical accident risk distribution probability of region  $v_i$  in history observed data,  $a(\Delta t)$  and  $b(\Delta t)$  are two coefficient indicating the global accident risk distribution in interval  $\Delta t$ .

$$p_i = \frac{\sum_{\Delta t} R(\Delta t, v_i)}{\sum_{\Delta t} \sum_j R(\Delta t, v_j)} \quad (3)$$

$$a(\Delta t) = \frac{R_{\max}(\Delta t, v_i)}{\log \frac{p_{\max}}{p_{\min}}}, b(\Delta t) = -a(\Delta t) * \log p_{\max} \quad (4)$$

### 3.2 Gated Graph Convolutional Multi Task (GGCMT) Model

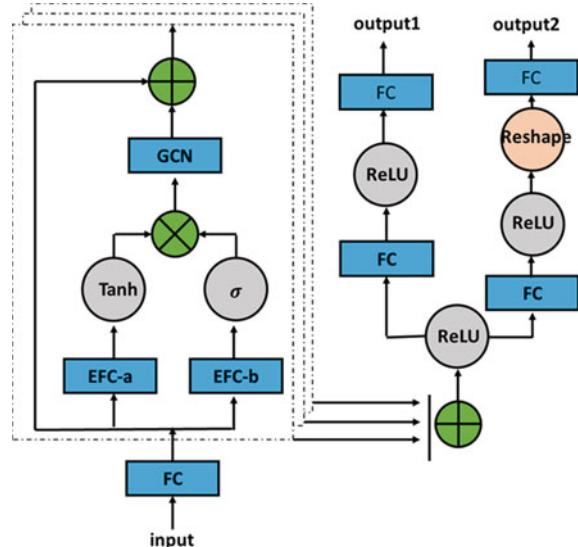
The framework overview of proposed gated graph convolutional multi task (GGCMT) model is illustrated in Fig. 3. The input data firstly goes through a fully connected layer for feature extraction, then the time series information will be analyzed through a gated structure network. And then the output of the gated structure will be fed into the graph convolutional neural network. Final outputs of the regional scale and global scale accident risk prediction values are calculated through two fully connected layers, respectively.

The gated structure network can be formulated as

$$X_{\text{gate}}^{l+1} = \text{Gate}_l(X^l) = \tanh(X^l U^l + b^l) \odot \sigma(X^l V^l + c^l) \quad (5)$$

where  $X^l \in R^{N \times n^{l-1}}$  is the input of the layer  $\text{Gate}_l$ ,  $X_g^{l+1} \in R^{N \times n^l}$  is the output.  $U^l, V^l \in R^{n^{l-1} \times n^l}$ ,  $b^l, c^l \in R^{n^l}$ , are learnable parameters.  $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  is the hyperbolic tangent function, value is between  $-1$  and  $1$ .  $\sigma(x) = \frac{1}{e^x + e^{-x}}$  is the

**Fig. 3** Framework overview of gated graph convolutional multi task (GGCMT) model



sigmoid function,  $\sigma(x) \in (0, 1)$ . And the operation  $\odot$  is the element-wise product, the extended fully connected layer calculated as  $EFC(X^l) = X^l U^l + b^l$ .

Taking advantage of its potential in non-Euclidean correlation and node risk propagation modeling [17], we use GCN to further analyze the output features of gated structure layer. The  $k$ -th GCN layer could be formulated as:

$$X_{GCN}^{k+1} = \text{GCN}_k(X^k) = \text{ReLU}((HX^k) * \theta^k) \quad (6)$$

where  $\theta^k \in R^{N \times m^{k-1} \times m^k}$  is the learnable parameter in the  $k$ -th layer,  $X_{GCN}^{k+1} \in R^{N \times m^k}$  and  $X^k \in R^{N \times m^{k-1}}$  are the output and input, respectively.  $\text{ReLU}(x) = \max(0, x)$  is the activation function. The operation symbol  $*$  indicates multidimensional array operator. For arrays  $A \in R^{d_1 \times d_2}$ ,  $B \in R^{d_1 \times d_2 \times d_3}$ , the operation  $C = A * B \in R^{d_1 \times d_3}$  satisfies that

$$c_{ij} = \sum_k^{d_2} a_{ik} b_{ikj}, \text{ where } 0 \leq i < d_1 \text{ and } 0 \leq j < d_3 \quad (7)$$

The matrix  $H$  denotes the Laplacian matrix of graph, which can be derived from the adjacency matrix  $W$ .

$$H = D^{-1} W \quad (8)$$

$$W = W_n + \lambda W_m \quad (9)$$

where  $W_n$  and  $W_m$  are weight matrices derive from static road network characteristics and urban road network image, respectively,  $\lambda = 0.2$  is a constant coefficient. In order to reduce the amount of computation, we limit the connectivity of the graph to  $\rho = 0.1$ , and ensure the sparsity of the affinity matrix of the graph matrix  $W$ .  $D \in R^{N \times N}$  is the diagonal matrix with  $D_{ii} = \sum_j w_{ij}$ .

Note that the role of reshape layer in the model is to flatten the dimension of the node which is used to calculate the overall risk accident output.

## 4 Empirical Studies

In this section, extensive empirical studies are carried out to evaluate our prediction framework on NYC Opendata dataset from January 1, 2017 to May 31, 2017. In the experiment, we select 70%, 20% and 10% datasets for training, evaluation and verification.

The length of time interval is set to be 30 min, and 48 intervals (24 h) are used to predict 6 (3 h) coming intervals accident risk. And the framework is evaluated by accuracy of top q (Acc@q) [18], which is widely adopted in spatiotemporal ranking

tasks, indicating the percentage of accurate predictions in nodes within  $k$  highest risks.

Three baselines are chosen for comparison, which are (1) ARIMA [19], a classic machine learning algorithm for time-series predictions; (2) SDAE [14], a model for real-time risk prediction incorporating human mobilities; (3) SDCAE [20], a method for citywide accident risk prediction.

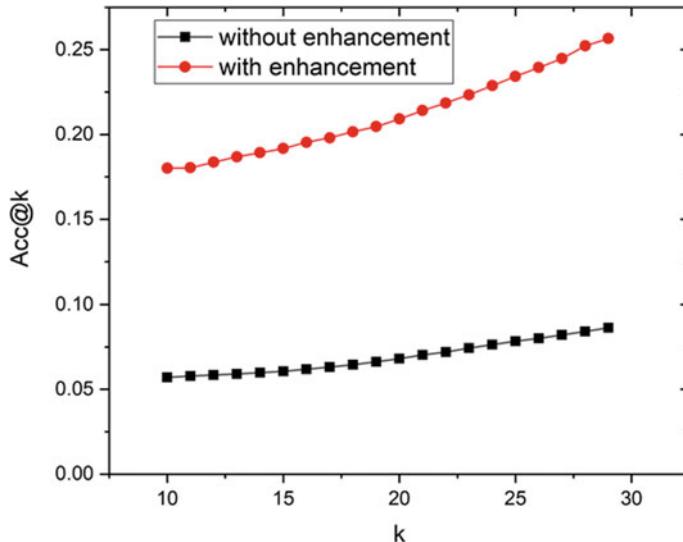
Table 1 illustrates the performance comparisons of our model with the baselines. Encouragingly, our framework outperforms the baseline models, having 6% higher accuracy than that of the baseline models.

Figure 4 shows the comparison of predicting results with (red curve) and without (black curve) the prior risk data enhancement method. Without data enhancement, the model tends to be partial to the most common but uninteresting cases without accident occurrence, whose risk value is 0. With the blessing of the enhancement, the top 20 accuracy of the model is improved from 6.81 to 20.98%, increasing by 14.17%.

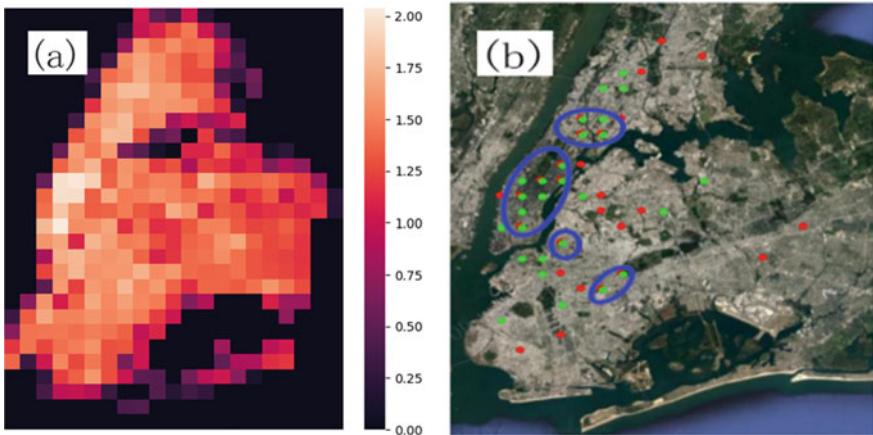
A case study is shown in Fig. 5. The heat map of accident risk prediction is presented in Fig. 5a, while the top  $k$  ( $k$  equals to the number of accidents in real situation) predicted nodes (green points) and the ground truth (red points) are shown

**Table 1** Performance comparisons of GGCMT and baselines

Models	ARIMA	SDAE	SDCAE	GGCMT
Acc@20	14.23%	12.08%	14.79%	20.98%



**Fig. 4** Comparison of predicting results with and without the prior risk data enhancement method



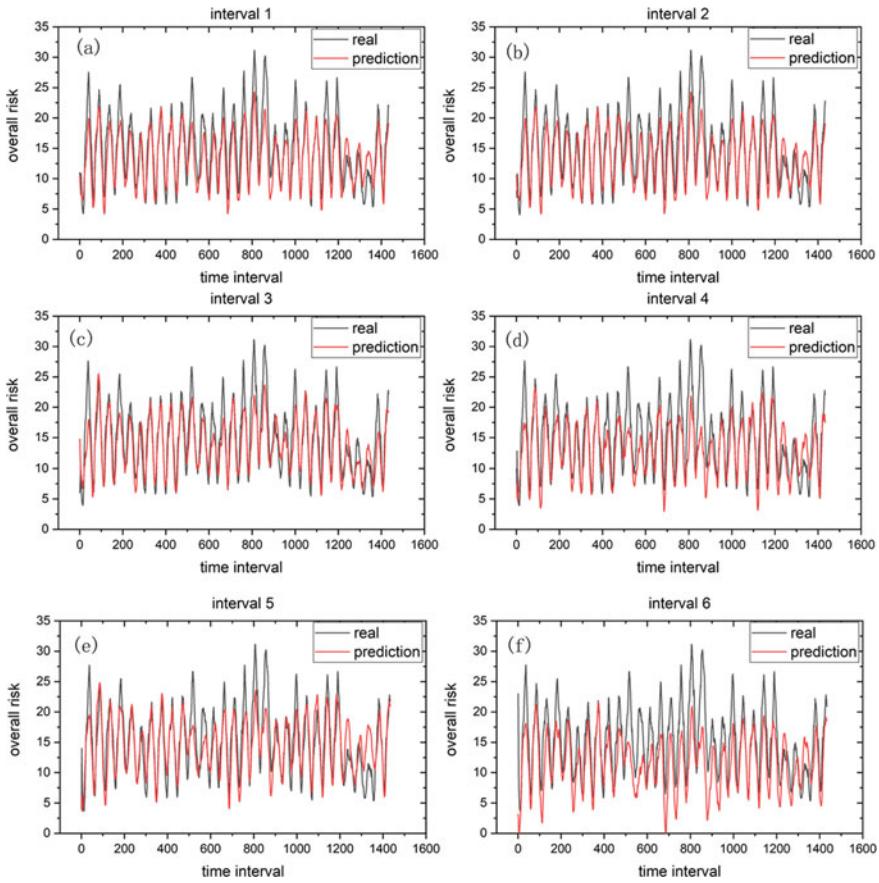
**Fig. 5** Case study results, **a** heat map of prediction, **b** accident distribution of predicted results and the actual situation

in Fig. 5b. One can see that a considerable number of traffic accidents have been accurately predicted.

In the auxiliary task, the overall accident risk is predicted to describe the development trend of accident risk in urban scale. The comparison of predicted overall risk and ground truth is presented in Fig. 6. The black curves present the real overall risks while the red curves present the prediction. As presented, the overall accident risk prediction is consistent with the real overall trend of traffic accidents well.

## 5 Conclusions

In this paper, (i) we adopt a computer vision method to analyze the urban road network image, and calculate visual similarities between different regions in urban city, since the visual similarities could reflect the similarity and correlation between different regions in urban road network space. (ii) To deal with the issue of sparse accident data, we propose a prior risk data enhancement method by replacing the zero values with indexes, indicating the historical traffic accident risk probability of nodes and the overall accident risk of the current moment. (iii) Furthermore, a gated graph convolutional multi task (GGCMT) model is proposed to predict the accident risk of both node scale and city scale. Results of empirical studies show that our framework outperforms baselines by 6% in top 20 accuracy, and the prior risk data enhancement method improve the accuracy by 14.17%. Besides, the overall accident risk prediction in the auxiliary task is in good agreement with the real overall risk.



**Fig. 6** Comparison of prediction (red curves) and real data (black curves) of overall risk. In order to better show the comparison results, we smooth both of the curves by exponential moving average method with weighted value 0.9

## References

1. Wu, Y., Tan, H.: Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework. arXiv preprint [arXiv:1612.01022](https://arxiv.org/abs/1612.01022) (2016)
2. Bai, C., Peng, Z.R., Lu, Q.C., et al.: Dynamic bus travel time prediction models on road with multiple bus routes. Comput. Intell. Neurosci. (2015)
3. Toqué, F., Côme, E., El Mahrbi, M. K., et al.: Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 1071–1076. IEEE (2016)
4. Usman, M., Carie, A., Marapelli, B., et al.: A human-in-the-loop probabilistic CNN-fuzzy logic framework for accident prediction in vehicular networks. IEEE Sens. J. (2020)
5. Zheng, M., Li, T., Zhu, R., et al.: Traffic accident's severity prediction: a deep-learning approach-based CNN network. IEEE Access **7**, 39897–39910 (2019)

6. Li, P., Abdel-Aty, M., Yuan, J.: Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Anal. Prevention* **135**, 105371 (2020)
7. Zhao, H., Cheng, H., Mao, T., et al.: Research on traffic accident prediction model based on convolutional neural networks in VANET. In: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 79–84. IEEE (2019)
8. Bogaerts, T., Masegosa, A.D., Angarita-Zapata, J.S., et al.: A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data. *Transp. Res. C Emerg. Technol.* **112**, 62–77 (2020)
9. Zhang, Y., Cheng, T.: Graph deep learning model for network-based predictive hotspot mapping of sparse spatio-temporal events. *Comput. Environ. Urban Syst.* **79**, 101403 (2020)
10. Wu, Z., Pan, S., Long, G., et al.: Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint [arXiv:1906.00121](https://arxiv.org/abs/1906.00121) (2019)
11. Zhou, Z., Wang, Y., Xie, X., et al.: RiskOracle: a minute-level citywide traffic accident forecasting framework. *Proc. AAAI Conf. Art. Intell.* **34**(01), 1258–1265 (2020)
12. Zhou, Z., Wang, Y., Xie, X., et al.: Foresee urban sparse traffic accidents: a spatio-temporal multi-granularity perspective. *IEEE Trans. Knowl. Data Eng.* (2020)
13. Wang, B., Luo, X., Zhang, F., et al.: Graph-based deep modeling and real time forecasting of sparse spatio-temporal data. arXiv preprint [arXiv:1804.00684](https://arxiv.org/abs/1804.00684) (2018)
14. Chen, Q., Song, X., Yamada, H., et al.: Learning deep representation from big and heterogeneous data for traffic accident inference. *Proc. AAAI Conf. Artif. Intell.* **30**(1) (2016)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
16. Koutini, K., Eghbal-Zadeh, H., Dorfer, M., et al.: The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification. In: 2019 27th European Signal Processing Conference (EUSIPCO), pp. 1–5. IEEE (2019)
17. Geng, X., Li, Y., Wang, L., et al.: Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. *Proc. AAAI Conf. Artif. Intell.* **33**(01), 3656–3663 (2019)
18. Liao, D., Liu, W., Zhong, Y., et al.: Predicting activity and location with multi-task context aware recurrent neural network. *IJCAI*, 3435–3441 (2018)
19. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50**, 159–175 (2003)
20. Chen, C., Fan, X., Zheng, C., et al.: SDCAE: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In: 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD), pp. 328–333. IEEE (2018)

# Exposing Video Frame Removal via Deep Features



Tianle Wu , Chunhui Feng , and Yigong Huang

**Abstract** In recent years, with the rapid increase of video data, its security problems have gradually been exposed, and people have begun to pay attention to it. Among them, video frame deletion is a common video forgery attack. In this paper, we propose a video frame deletion detection method using deep features. We first use the deep features of AlexNet pre-trained on the ImageNet dataset to calculate the inter-frame distortion level. After that, a feature enhancement method based on the local average distortion level is proposed to eliminate the fluctuation of the video content and obtain the relative distortion level. Finally, the generalized ESD test is used to detect the frame deletion points in the video. We created two datasets containing large-scale complex content scenarios for testing the proposed method. The experimental results show that the F1 score of the proposed method reaches 90.26%, which is significantly better than the existing methods on the created dataset.

**Keywords** Video forensics · Frame deletion · Deep features · Distortion level · Relative distortion level

## 1 Introduction

Nowadays, modifying digital video content is effortless. People can use editing software on a computer or smartphone to edit the video at will. In many countries, digital video has a strong legal effect. Therefore, verifying the authenticity and integrity is a necessary condition for digital video as evidence in court. Besides, the authenticity and integrity of digital video are also vital in news, security and other fields.

Video frame deletion is to delete several consecutive frames from the video. After careful frame deletion, it is difficult to be noticed by humans. Therefore, it is challenging to detect and locate the frame deletion operation in the video. In [1], Wang et al. found that frame deletion will cause double compression of I frames and increase motion estimation. They use this phenomenon to detect video frame

---

T. Wu · C. Feng ( ) · Y. Huang

College of Computer and Information Sciences, Fujian Agriculture and Forestry University,  
Fuzhou 350002, China

e-mail: [fengchunhui@fafu.edu.cn](mailto:fengchunhui@fafu.edu.cn)

deletion. However, when the number of frame deletions is an integer multiple of GOP, their method is invalid. Stamm [2] used the prediction error of P frames to design a method that can detect the deletion of video frames without being restricted by the GOP size. Many studies have explored the characteristics of compressed domains, such as [3, 4]. Although the compressed domain feature can effectively detect the deletion of video frames, it is limited by the video compression format. Therefore, Wang et al. [5] started the exploration in the spatial domain. They found that in interlaced video, the motion between a single field and adjacent fields is equal. They use this phenomenon to detect the forgery of interlaced video. After that, because the spatial domain features do not restrict the compression format, many researchers focus on exploring spatial domain features. Chao et al. [6] studied the changes of optical flow in the video and found that inter-frame forgery would affect optical flow consistency. Feng et al. [7] analyzed the motion of different intensities in the video and proposed a motion-adaptive method to detect video frame deletion. There are many ways to use spatial domain features, such as velocity field intensity [8], ENF [9], etc. However, these methods using spatial domain features are easily affected by the content of the video, such as lighting changes, motion speeds, and their robustness is poor.

To solve the problems mentioned above, we propose a method to effectively detect the deletion of video frames. We first use the Learned Perceptual Image Patch Similarity (LPIPS) to measure the inter-frame distortion level. For the distance between the two images, the LPIPS can be consistent with human visual judgment. Then we use a feature enhancement algorithm based on the local average distortion level to eliminate the effects of lighting changes, motions of different intensities, and other video content. Finally, we use the generalized ESD test to detect inter-frame abnormalities in the video. We tested the proposed method on two datasets with complex content scenarios and compared it with existing methods. The results show that our method has a good performance.

## 2 Review of Learned Perceptual Image Patch Similarity

Learned Perceptual Image Patch Similarity (LPIPS) is a new image consistency metric proposed by Zhang et al. [10], which is used to judge the perceptual distance of two images. Unlike traditional pixel-by-pixel image similarity measurements such as structural similarity (SSIM) or L2 Euclidean distance, LPIPS uses deep neural network features to calculate the perceived distance between two images. It can be obtained by training different network architectures at different tasks and supervision levels.

The calculation process of the LPIPS of the two images is shown in formula (1). First, a network  $F$  is given, and then the feature stack of the  $L$  layer in the network is extracted and unit-normalized in the channel dimension. Then use a vector  $w$  to scale each channel and calculate the L2 distance. Finally, the spatial dimension and all layers are averaged.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l)\|_2^2 \quad (1)$$

where  $x, x_0$  are the two image patches of the input network.  $\hat{y}^l, \hat{y}_0^l \in \mathbb{R}^{H_l \times W_l \times C_l}$  is the feature of the  $l$  layer;  $w_l \in \mathbb{R}^{C_l}$ .

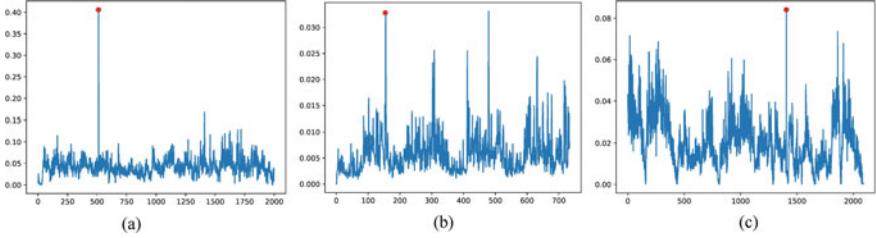
### 3 Proposed Method

In this section, we first discuss the basic principle of using LPIPS to expose frame deletion points (FDPs) in a video. Then, a feature enhancement algorithm to obtain the relative distortion level is introduced to eliminate the influence of video content. Finally, the method of detecting FDPs from the relative distortion level sequence of the video is introduced, and our overall detection framework is given.

#### 3.1 LPIPS of Frame Deleted Video

In [10], Zhang et al. proved that LPIPS has the comparable ability as the human vision in distinguishing the original image and different types of distortion images (traditional distortion and CNN-based distortion). In this article, we regard the process from the current frame to the next frame of a video as a special kind of distortion (that is, the current frame is the original image, and the next frame is the distorted image). This assumption is based on the characteristics of the video itself. In the video, the shooting interval of adjacent frames is very short, and the content changes slightly and is very similar. This characteristic also determines that the distortion between adjacent frames is not a single distortion but a complicated distortion that mixes various single distortions such as lighting changes, movement, heavy Shadows, etc.

In order to observe the changing law of the distortion level in the frame deletion video, we visualized the distortion levels of each inter-frame of three deleted frame videos with different motion intensities, as shown in Fig. 1. Due to the influence of the video content, the changes in inter-frame distortion level are proportional to the level of change in video content. In a video, the distortion level between adjacent frames that have not undergone the frame deletion operation is roughly the same. After the frame deletion operation, the distortion level between adjacent frames will increase due to the gap of video content and destroy the continuity of the distortion level of the entire video. Therefore, we can use the LPIPS consistency of the video inter-frame to detect whether there is a frame deletion operation in the video (that is, to detect inter-frame with a sudden increase in distortion level).



**Fig. 1** The inter-frame distortion level curves of three frame-deleted videos with different motion intensities, where the red dot marks the FDP. The movement speed of the three videos is **a** > **c** > **b**

### 3.2 Feature Enhancement Algorithm

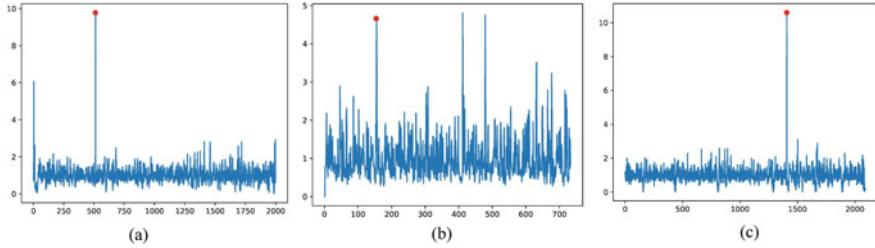
In the previous section, we analyzed the characteristics of the distortion level between the (non-deleted frames point) non-FDP and the FDP. We found that the inter-frame distortion level fluctuates with the video content, and a small part of the fluctuation causes the distortion level to be like FDP. The saliency of the FDP distortion level in the entire video will be severely affected. After observation, we found that the increase in distortion level caused by frame deletion is sudden, while the distortion level caused by content fluctuations is continuous. Therefore, to make the distortion level of the FDP more significant, we designed a feature enhancement algorithm according to the difference in the increase of the distortion level of the non-FDP and FDP. The algorithm is based on the average distortion level of past inter-frame and future inter-frame. Our feature enhancement algorithm is shown in Eq. (2). For the  $i$ -th inter-frame of a video, we first calculate the average distortion level of the past and future  $2w$  inter-frames. Then use the quotient of the distortion level of  $i$ -th inter-frame and the average distortion level to expose the sudden increase in the distortion level and eliminate the continuous increase.

$$RLPIPS(i)$$

$$= \begin{cases} \frac{LPIPS(i)}{\frac{1}{i-1+w} \left( \sum_{j=0}^{i-1} LPIPS(j) + \sum_{j=i+1}^{i+w} LPIPS(j) \right)}, & i - w < 0 \\ \frac{LPIPS(i)}{\frac{1}{2w} \left( \sum_{j=i-w}^{i-1} LPIPS(j) + \sum_{j=i+1}^{i+w} LPIPS(j) \right)}, & i - w \geq 0 \text{ and } i + w \leq N \\ \frac{LPIPS(i)}{\frac{1}{w+N-i-1} \left( \sum_{j=i-w}^{i-1} LPIPS(j) + \sum_{j=i+1}^N LPIPS(j) \right)}, & i + w > N \end{cases} \quad (2)$$

where  $LPIPS(i) = d(i, i + 1)$  is the distortion level from the  $i$ -th frame to the  $i + 1$ -th frame in the video.  $N$  is the number of frames of the video.

After feature enhancement, we call the obtained sequence the relative distortion level sequence. When  $w = 15$ , the feature enhancement results of the three videos in Fig. 1 are shown in Fig. 2. The relative distortion level of the FDP is significant in the entire video. The relative distortion level of the non-FDP in the video is very stable and is no longer affected by the fluctuation of the video content.



**Fig. 2** The inter-frame relative distortion level curve of three frame-deleted video. The red dot is the FDP

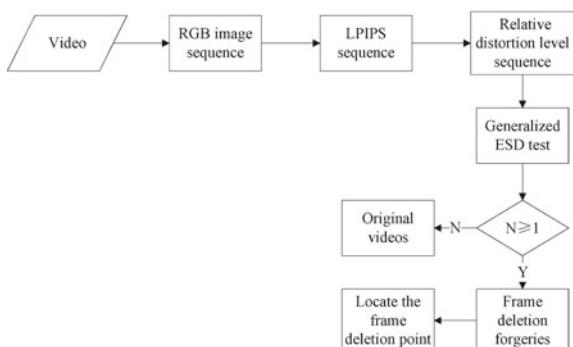
### 3.3 Frame Deletion Point Detection

The generalized ESD test [11] is an anomaly detection method that can detect abnormal points in normally distributed data. In this paper, we use the relative distortion level to expose the FDPs in the deleted frame video. The relative distortion level of the FDPs breaks the continuity of the overall relative distortion level of the video and presents abnormal discontinuities. Besides, the relative distortion level sequence data is approximately normally distributed. Therefore, we can use the generalized ESD test to detect the abnormal points in the relative distortion level sequence to determine the FDP in the video. There are two critical parameters in the generalized ESD test: the number of abnormal points  $r$  and the significance level  $\alpha$ . In our method, we set  $r = 1$  and  $\alpha = 0.05$ .

### 3.4 Overall Framework of the Proposed Method

The overall flow of our detection algorithm is shown in Fig. 3. Given a video, we first decode the video to an RGB image sequence and resize the image to  $64 \times 64$  pixels. Then use the 5-layer convolution feature in the pre-trained Alex network to calculate

**Fig. 3** Flow chart of video frame deletion detection algorithm



the LPIPS between each inter-frame to obtain the LPIPS sequence of the entire video. Linear correction and spatial averaging are used in the calculation of LPIPS. Then use the feature enhancement algorithm to process the LPIPS sequence to obtain the relative distortion level sequence. Finally, the generalized ESD test is used to detect the abnormal points in the relative distortion level sequence and obtain the number of abnormal points  $N$  and their index in the relative distortion level sequence. If the number of abnormal points  $N < 1$ , the video is determined to be the original video. Otherwise, it is the frame deleted video. Besides, if a video is judged to be a deleted frame video, the abnormal point index is the position in the video where the FDP is located.

## 4 Experiment

### 4.1 Dataset

In our experiments, we evaluate the proposed method using two video datasets of multiple scenes with different shooting methods. The original videos of the two datasets are from VISION [12] and VFDD2.1 [13]. These two datasets are obtained by editing the original video.

We use 798 original videos of about 1 min and 10 s in VISION to create our D-VISION dataset. Create our D-VFDD dataset using 714 original videos with a length of about 10 s in VFDD2.1. Our dataset is generated to delete 1 s-length frames from random locations in the original video.

### 4.2 Evaluation Metrics

We use the recall, precision, and F1 score shown in formulas (3)–(5) to evaluate our method.

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (5)$$

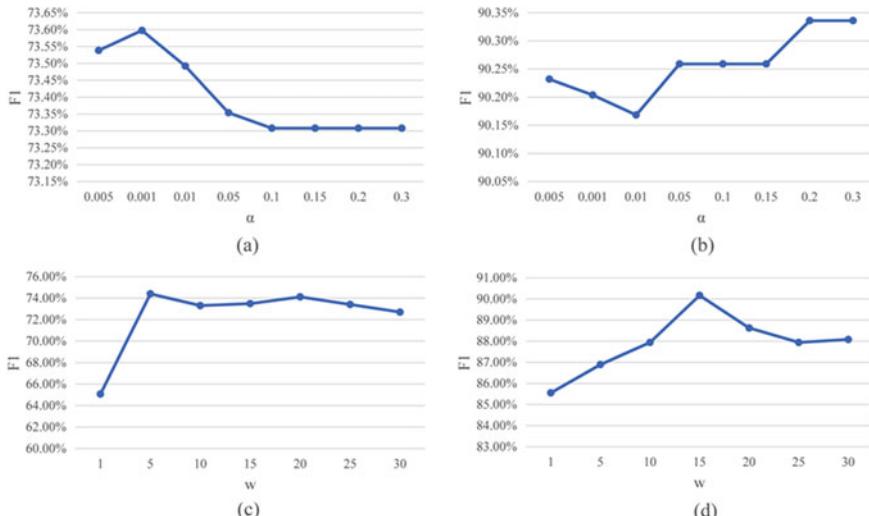
where TP (True Positive) and TN (True Negative) are the numbers of correctly recognized FDPs and non-FDPs. FP (False Positive) is the number of non-FDPs falsely

classified as FDPs. FN (False Negative) is the number of FDPs falsely recognized as non-FDP frames.

### 4.3 Discussion of Parameters

In this section, we discuss the window size  $w$  involved in the proposed feature enhancement algorithm and the significance level  $\alpha$  in anomaly detection. Figure 4a and b show how the F1 score changes with  $\alpha$  when  $w = 15$ . We can see that as  $\alpha$  increases, the F1 score shows an opposite trend on the D-VISION dataset and the D-VFDD dataset. The possible reason is that the number of video frames in the D-VISION data set is much larger than that in the D-VFDD dataset. As  $\alpha$  increases, the stringency of the generalized ESD test decreases, causing more false alarms for videos with more frames. Therefore, to make our method have satisfactory generalization. We set the parameter  $\alpha$  to 0.05.

Figure 4c, d shows the effect of the window size  $w$  in the feature enhancement algorithm on the F1 score when  $\alpha$  is 0.01. We can see that as  $w$  increases, F1 scores show roughly the same trend on the two datasets (rising first and then falling). Therefore, we set the window size  $w = 15$  in the feature enhancement algorithm.



**Fig. 4** The effect of different  $w$  and  $\alpha$  on the performance of the algorithm. Among them, **a, c** are the results on the D-VISION dataset, and **b, d** are the results on the D-VFDD dataset

**Table 1** Compared with the performance of existing methods

	D-VISION		F1 (%)	D-VFDD		
	P (%)	R (%)		P (%)	R (%)	
HOG [15]	8.70	68.69	15.45	20.85	85.71	33.54
LBP [14]	13.35	45.45	20.64	83.74	76.47	79.94
LPIPS	69.01	68.92	68.96	81.10	78.71	79.89
Ours	73.40	73.31	73.35	90.32	90.20	90.26

#### 4.4 Performance Comparison with State-of-the-Art Methods

We tested the performance of the proposed method to detect video frame deletion on the D-VISION dataset and D-VFDD dataset and compared them with representative frame deletion detection algorithm [14] and [15]. For fairness, we chose the best parameters recommended in the two methods during the test. Table 1 shows the performance of the three methods. We can see that the F1 score of the proposed method exceeds the other two methods on both datasets. The existing best method has 50.72% lower detection performance than our method on the D-VISION dataset and 10.32% lower detection performance than our method on the D-VFDD dataset. It can be seen that the performance of the three methods on the D-VISION dataset is lower than that of the D-VFDD dataset. There are two possible reasons. The first reason is that the video in the D-VISION dataset is longer than the video in D-VFDD. Longer videos will generate more false positives. The second reason is the flat video in the D-VISION dataset. The videos belonging to flat surfaces include walls and skies. Especially for wall video, the wall content is single and lacks effective feature points, which makes non-FDP and FPD extremely similar.

## 5 Conclusion

In this paper, we propose a novel method to detect video frame deletion. We assume that the change of adjacent frames of the video is a kind of complex distortion and reveals the FDP according to the level of distortion. The distortion level is calculated by the deep features extracted from the trained deep neural network. The relative distortion level is proposed, which effectively improves the influence of video content fluctuation. Use generalized ESD to detect FDP in the video. The experimental results prove the superiority of the proposed method. Nevertheless, the detection effect of video with complex and changeable content scenes still needs to be improved. In the future, we will further study the proposed method. Explore the complex transformation relationship between adjacent frames of video and build a distortion model between adjacent frames of video based on deep learning.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China under Grant 61802064.

## References

1. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting double MPEG compression. In: Proceedings of the 8th Workshop on Multimedia and Security, pp. 37–47 (2006)
2. Stamm, M.C., Lin, W.S., Liu, K.J.R.: Temporal forensics and anti-forensics for motion compensated video. *IEEE Trans. Inf. Forensics Secur.* (2012)
3. Yu, L., Wang, H., Han, Q., et al.: Exposing frame deletion by detecting abrupt changes in video streams. *Neurocomputing* **205**, 84–91 (2016)
4. Hong, J.H., Yang, Y., Oh, B.T.: Detection of frame deletion in HEVC-coded video in the compressed domain. *Digit. Investig.* **30**, 23–31 (2019)
5. Wang, W., Farid, H.: Exposing digital forgeries in interlaced and deinterlaced video. *IEEE Trans. Inf. Forensics Secur.* **2**, 438–449 (2007)
6. Chao, J., Jiang, X., Sun, T.: A Novel Video Inter-frame Forgery Model Detection Scheme Based on Optical Flow Consistency. *Lecture Notes in Computer Science*, pp. 267–281 (2013)
7. Feng, C., Xu, Z., Jia, S., et al.: Motion-adaptive frame deletion detection for digital video forensics. *IEEE Trans. Circuits Syst. Video Technol.* **27**, 2543–2554 (2017)
8. Wu, Y., Jiang, X., Sun, T., Wang, W.: Exposing video inter-frame forgery based on velocity field consistency. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2674–2678. IEEE (2014)
9. Wang, Y., Hu, Y., Liew, A.W.-C., Li, C.-T.: ENF based video forgery detection algorithm. *Int. J. Digit. Crime Forensics* **12**, 131–156 (2020)
10. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
11. Rosner, B.: Percentage points for a generalized ESD many-outlier procedure. *Technometrics* **25**, 165 (1983)
12. Shullani, D., Fontani, M., Iuliani, M., et al.: VISION: a video and image dataset for source identification. *EURASIP J. Inf. Secur.*, 15 (2017)
13. Li, J.C., Hu, Y.J., Mohammed, A.A., et al.: Expansion of video forgery detection database and validation of its effectiveness. *Yingyong Kexue Xuebao/J. Appl. Sci.* **36**, 347–361 (2018)
14. Zhang, Z., Hou, J., Ma, Q., Li, Z.: Efficient video frame insertion and deletion detection based on inconsistency of correlations between local binary pattern coded frames. *Secur. Commun. Networks* **8**, 311–320 (2015)
15. Fadl, S., Han, Q., Qiong, L.: Exposing video inter-frame forgery via histogram of oriented gradients and motion energy image. *Multidimens. Syst. Signal Process.* **31**, 1365–1384 (2020)

# Research on Semantic Segmentation and Object Grasping Strategy Generation Based on Deeplab Algorithm



Shaobo Li<sup>✉</sup>, Qiang Bai<sup>✉</sup>, Jing Yang<sup>✉</sup>, Liya Yu<sup>✉</sup>, and Guangwei Wang<sup>✉</sup>

**Abstract** The Semantic segmentation is another important research direction in the field of machine vision after object recognition. And the contour segmentation of objects will be realized at the pixel level, which will provide great help for robot object grasp strategy generation and driverless driving. The algorithm of deeplab series proposed by Google innovatively use dilated convolution kernels instead of traditional solid convolution kernels to achieve precise object semantic segmentation and perform well on multi-scale objects. Firstly, this paper studies the principle of deeplab algorithm from a mathematical point of view, and discusses its excellent performance in semantic segmentation; Secondly, this paper applies the advantages of deeplab algorithm in semantic segmentation to the research of object grasp strategy generation, which makes a beneficial exploration for robot grasp and promotes the wider application of robot; Finally, the core parameters of deeplab algorithm are optimized, and the experimental results are analyzed in detail to improve the accuracy of the model.

**Keywords** Semantic segmentation · Grasping strategy · Deeplab

## 1 Introduction

Artificial intelligence (AI) is considered to be the fourth industrial revolution, and the world's top technology companies (Google, Facebook, Alibaba, Tencent, Baidu, etc.) all regard AI as an important strategic research direction. AI technology represented by deep learning has achieved world-renowned achievements in the fields of object recognition, speech translation, natural language processing, and driverless driving. In the field of machine vision, deep learning has achieved excellent performances in image recognition, object location, and semantic segmentation. Semantic segmentation solves the above two problems from the pixel level and it is also a popular research direction in the field of computer vision. Fully Convolutional Networks (FCN) is

---

S. Li · Q. Bai (✉) · J. Yang · L. Yu · G. Wang  
Guizhou University, Guiyang 550000, Guizhou, China  
e-mail: [cme.qbai18@gzu.edu.cn](mailto:cme.qbai18@gzu.edu.cn)

an early classic algorithm in the field of semantic segmentation, and it provides a basic research framework for semantic segmentation [1]. It has the advantages of accepting input images of any size and high computational efficiency, however, the image details are not sensitive and the model training process is cumbersome and the connection between pixels cannot be fully considered. The U-Net network [2] proposed by Olaf Ronneberger et al. in 2015 is a variant of FCN, which implements image feature extraction and semantic segmentation based on the Encoder-Decoder structure. Compared with FCN, U-Net only needs to be trained once to achieve model training, and it has achieved better semantic segmentation performance in the field of medical images. In 2015, Yu et al. [3] proposed a semantic segmentation algorithm based on atrous-convolutions. This method expands the convolution system without losing resolution and realizes the perception of multi-scale context information, but it has the disadvantages of large amount of calculation and large memory requirements. In the field of semantic segmentation, the algorithms of deeplab series have achieved world-renowned achievements, and they have become more and more perfect after multiple iterations. In 2014, the deeplabV1 [4] effectively solved the problem of insufficient accuracy in the semantic segmentation of the DCNNs, and made up for the disadvantage of the advanced feature translation invariance of DCNNs. Although DeeplabV1 achieves end-to-end training, and solves the problem of resolution degradation and partial loss of image details caused by downsampling, the main focus of this model is the fully connected CRFs used for subsequent positioning. Aiming at the problem of the difficulty of multi-scale object recognition and the local accuracy degradation caused by the inherent invariance of DCNNs, Liang-Chieh Chen et al. proposed deeplabV2 [5] in 2016, adding a multi-field of view on the basis of V1 and introducing the space pyramid model—Atrous Spatial Pyramid Pooling (ASPP), and the segmentation problem of different size objects are effectively solved. In addition, the base structure of the model is replaced by ResNet from VGG16, which can achieve deeper model training. The DeeplabV3 algorithm proposed in 2017 [6] improved the ASPP structure, and spliced the expansion convolution results of different expansion rates (cascaded multiple atrous-convolutions structures). Different from dilated convolutions [3], V3 directly expands the convolution of the feature map in the middle, and at the same time the conditional random field (CRF) is deleted. The whole model is concise and easy to understand, but it also has the disadvantages of poor output pictures and too little information. The deeplabV3+ algorithm [7] proposed in 2018 modified the backbone of the model to modify xception [8] and used the depth-wise separable convolution structure in the ASPP and decoding modules to improve the accuracy and speed of the image segmentation algorithm, while protecting the edge details of the object information, but due to the model structure is more complex, the running speed is slower than other algorithms. Based on the excellent performance of deeplabV3+ algorithm in the field of semantic segmentation, this paper applies it to the research of robot grasping strategy generation.

The article is divided into four parts: The first part elaborates on the advantages and disadvantages of the current mainstream semantic segmentation algorithms; The second part analyzes the excellent performance of deeplab algorithm from the

perspective of mathematical theory; The third part mainly talks about the improvement of the deeplab algorithm and apply in the generation of grasp strategy; The fourth part summarizes the content of the article and makes a conclusion.

## 2 Principle Analysis

Traditional deep convolutional neural networks use solid convolution kernels and pooling, which will cause the loss of internal data structure and spatial hierarchical information, and will also cause the information of small objects to be unable to be reconstructed. This has caused the research of semantic segmentation to be in a bottleneck period and cannot achieve a significant improvement in accuracy, and the design of dilated convolution avoids these problems well. However, how to properly design the structure of the dilated convolution will determine its performance in semantic segmentation. If only the kernel with the same dilation rate value is superimposed multiple times, or the discontinuous kernel means that not all pixels are used for calculation, so the continuity of the information will be lost by treating the information as a checker-board. In response to the above situation, the algorithms of deeplab series have been improved (as shown in Eqs. 1–3) to achieve the fusion of different dilated rate convolution kernels. The  $r$  value in the Eqs. 2 and 3 represents different dilated rates. The deeplabV3+ algorithm achieves accurate perception and semantic segmentation of multi-dimensional objects by fusing convolution kernels of different dilated rate and performing parallel operations.

$$\text{Ordinary convolution: } y = \omega x + b \quad (1)$$

$$\text{Dilated convolution: } y = (\omega + r)x + b \quad (2)$$

DeeplabV3+:

$$\begin{aligned}
 &y = \omega x + b \\
 &y = (\omega + r_1)x + b \\
 &y = (\omega + r_2)x + b \\
 &y = (\omega + r_3)x + b
 \end{aligned}
 \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \text{Fusion} \quad (3)$$

Image pooling

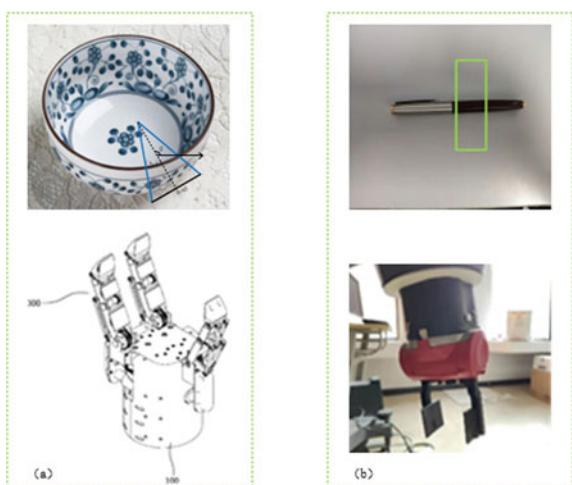
### 3 The Actual Effect of the Model on the Grasp Strategy

Vision is the main way for humans to receive all kinds of information and researchers hope that robots can have a vision system with high accuracy and robustness like human to help people complete all kinds of work. Therefore, machine vision has always been an important research topic in the field of AI and robot. With the rapid development of deep learning technology, it has been widely and successfully applied in defect detection, object recognition, medical image judgment and other fields [9–19]. However, researchers hope that machine vision can achieve the accurate generation of object grasp strategy and lay the foundation for object grasp, which requires high semantic segmentation accuracy.

#### 3.1 Idea

The deeplab series semantic segmentation algorithm proposed by Google in 2014 has achieved the best semantic segmentation performance on Pascal dataset (71.6% of mean IOU). After four iterations, the deeplabv3+ algorithm proposed in 2018 has become one of the best semantic segmentation algorithms and the precision of 89.0% and 82.1% has been achieved in Pascal VOC 2012 and cityscaps test set. Semantic segmentation is a pixel level image processing technology, it can achieve high-precision object segmentation, which lays a solid foundation for the generation of grasp strategy. As shown in Fig. 1, the current mainstream grasping methods are mainly divided into triangle grasping strategy (three fingers dexterous hand) and rectangular grasping strategy (two fingers dexterous hand). Two fingers dexterous hand has the advantages of low cost and simple structure, but its versatility is poor,

**Fig. 1** Mainstream grasping methods: **a** triangle grasping strategy; **b** rectangular grasping strategy



especially when grasping spherical or cylindrical objects, it has a high risk of slipping. Although the cost of three finger dexterous hand is slightly higher than that of the two finger dexterous hand, it is more similar to the human grip action, so it has higher versatility. Based on the above analysis, this paper will study the grasping strategy of three fingers dexterous hand based on deeplab semantic segmentation algorithm. As shown in Fig. 1a, the three fingers grasping strategy is based on a four-dimensional grasping representation with a fixed orientation triangle:  $(x, y)$  is the center coordinates of triangle, angle  $\theta$  represents the angle between the triangle and the horizontal direction, height  $\omega$  represents the height of a triangle with  $d$  as its base. Figure 1b shows the rectangular grasping strategy. At present, the mainstream rectangular grasp strategy mainly adopts the seven dimensional grasp representation method, which is similar to the method proposed by Morrison et al. [20]. The grasp attitude is expressed as follows:

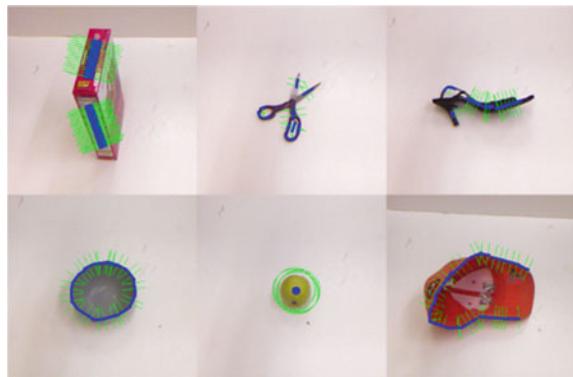
$$G_r = (P, \Theta_r, W_r, Q) \quad (4)$$

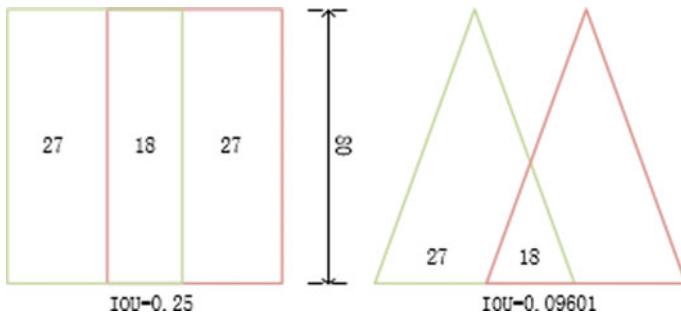
where  $P = (x, y, z)$  is the center coordinates of the end effector,  $\Theta_r$  is the rotation angle of the end effector around the Z axis,  $W_r$  is the required width of the end effector, and  $Q$  is the grasp strategy quality score.

### 3.2 Dataset

Cornell University's grasp dataset is the mainstream grasp dataset at present. The dataset includes 885 images of 240 objects, in which the label of the original data is rectangle grasp strategy. In order to realize the training of triangle grasp strategy model, this paper uses Dexin et al. [21] annotated dataset for model training. As shown in Fig. 2, the annotation method of the dataset is completely different from the rectangular grasp strategy. The points in the blue area are all graspable points, and a green line is drawn at the end of each grasp point to indicate the direction of

**Fig. 2** Triangle grasp strategy label [20]



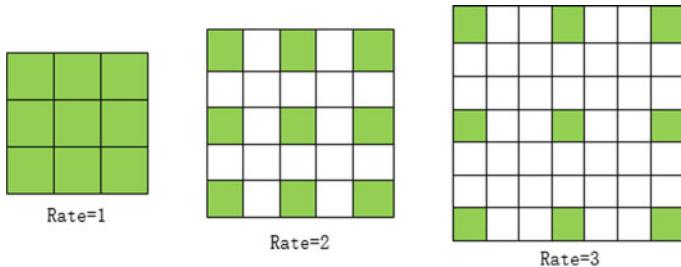


**Fig. 3** IOU values of different grasp representations

the middle line of the grasp triangle. The green line represents half of the grasping width. It is worth noting that the green line cannot be drawn symmetrically for the objects that cannot be grasped symmetrically and only the objects that can be grasped symmetrically can be drawn symmetrically. The green circle indicates that there is no limit on the grasp angle. As shown in Fig. 3, the rectangle representation is very different from the triangle representation, in which the green box represents the expected value and the red box represents the predicted value.

### 3.3 Training and Testing

One of the cores of deeplab series algorithm is the use of dilated convolution kernel with different dilated rates (as shown in Fig. 4), which makes convolution kernel have strong sense field and can control the resolution of the feature calculated by deep CNN. At the same time, combining the dilated convolutions with different parameters in series and parallel can also deal with the problem of semantic segmentation under multi-scale conditions. After reading a large number of papers, it is found that the receptive fields of convolution kernels with different dilated rate values are very



**Fig. 4** Convolution kernels with different rates

different [22–27]. Therefore, this paper will test the accuracy of the model with different dilated rate values, and make a useful exploration for object grasp.

### 3.4 Evaluation

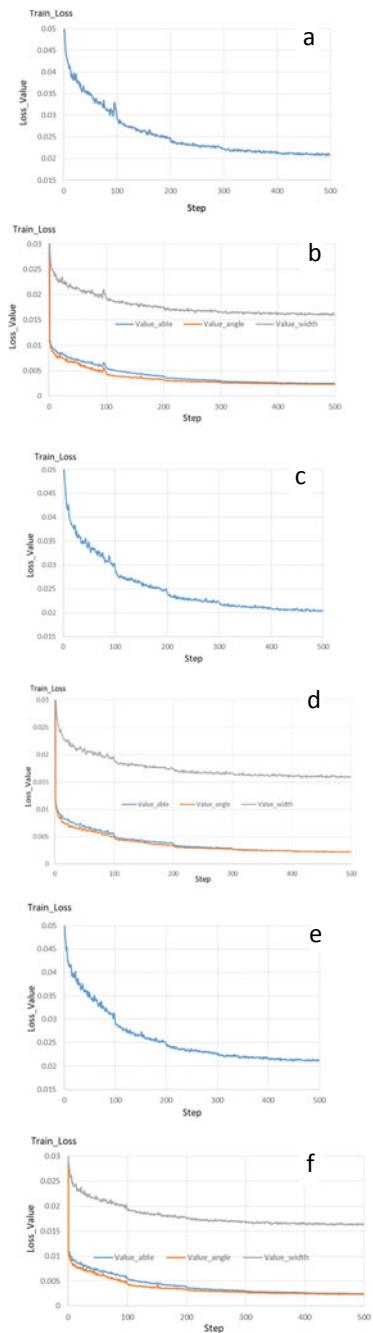
According to the existing research results [22–27], this paper designs three cases of differentiated dilated rate: [1, 3, 6, 9], [1, 6, 12, 18], [1, 3, 5, 7], and verify the performance differences of the model under different dilated rate. As shown in Fig. 5, the convergence of the model varies greatly under different dilated rate, especially in the first half of the model training (before 200 step), when the dilated rate is 1-3-6-9, the model has a more violent oscillation phenomenon, which may be caused by the small rate value and the common divisor, otherwise, the convergence of the model is improved when the discounted rate is set to 1-6-12-18 and 1-3-5-7.

In this paper, 92 images are randomly selected as the test set in the Cornell University grasp dataset and Table 1 describes the performance of the model in the test set under different dilated rates. It is found that the model has the best performance when the dilated rate is 1-6-12-18. This is due to the large span and uniform distribution of this group of parameters, so it can achieve better semantic segmentation effect, and then improve the accuracy of the grasp strategy generation.

## 4 Conclusion

Fast and reasonable grasping strategy is the premise of robot grasping, so it has important research value. Based on the excellent performance of deeplab algorithm in semantic segmentation, this paper uses it to generate object grasp strategy. Through theoretical derivation and analysis of existing research results, three groups of different dilated rates are selected as experimental variables. Through model training and test verification, it is found that the model can achieve best performance when the dilated rate is 1-6-12-18, which achieves 96.8% accuracy in the test set. In this paper, the model of high-precision grasping strategy is studied, which makes a useful exploration for the research of robot grasping.

**Fig. 5** The training results of the model under different dilated rates. **a** The overall loss function curve of the model with the dilated rate of 1-3-6-9; **b** The loss function curve of confidence value, angle and width with the dilated rate of 1-3-6-9; **c** The overall loss function curve of the model with the dilated rate of 1-6-12-18; **d** The loss function curve of confidence value, angle and width with the dilated rate of 1-6-12-18; **e** The overall loss function curve of the model with the dilated rate of 1-3-5-7; **f** The loss function curve of confidence value, angle and width with the dilated rate of 1-3-5-7



**Table 1** Performance statistics of models with different dilated rates

Dilated rate	Loss	Loss_able	Loss_angle	Loss_width	Test accuracy (%)
1-3-6-9	0.020863	0.002477	0.002351	0.016035	92.4
1-6-12-18	0.020863	0.002477	0.002351	0.016035	96.8
1-3-5-7	0.020863	0.002477	0.002351	0.016035	89.1

**Acknowledgements** This work was supported in part by the National key technologies R&D program of China (Grant No. 2018AAA0101800), the Cultivation project of Guizhou University (Grant No. [2019]22), the Talent introduction research project of Guizhou university (Grant No. [2020]14), the Science and Technology Foundation of Guizhou Province (Grant No. [2020]1Y233)

## References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. IEEE, Boston, MA, USA (2015)
2. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Cham, Springer (2015)
3. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico (2016)
4. Chen, L., et al.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA (2015)
5. Chen, L., et al.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2017)
6. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. ArXiv, abs/1706.05587 (2017)
7. Chen, L., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818. ECCV, Munich, Germany (2018)
8. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. IEEE, Honolulu, HI, USA (2017)
9. Bozhkov, L., Georgieva, P.: Overview of deep learning architectures for EEG-based brain imaging. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE, Rio de Janeiro, Brazil (2018)
10. Shen, X., et al.: Spatial-temporal human gesture recognition under degraded conditions using three-dimensional integral imaging: an overview. In: 2018 17th Workshop on Information Optics (WIO), pp. 1–3. IEEE, Québec, QC, Canada (2018)
11. Gite, B., Nikhal, K., Palnak, F.: Evaluating facial expressions in real time. In: 2017 Intelligent Systems Conference (IntelliSys), pp. 847–855. IEEE, London, UK (2017)
12. Panchal, P., Raman, V.C., Mantri, S.: Plant diseases detection and classification using machine learning models. In: 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) , pp. 1–6. IEEE, Bengaluru, India (2017)

13. Gao, M., et al.: RGB-D-based object recognition using multimodal convolutional neural networks: a survey. *IEEE Access* **7**, 43110–43136 (2019)
14. Wang, H., et al.: A comprehensive overview of person re-identification approaches. *IEEE Access* **8**, 45556–45583 (2020)
15. Celebi, M.E., Codella, N., Halpern, A.: Dermoscopy image analysis: overview and future directions. *IEEE J. Biomed. Health Inf.* **23**(2), 474–478 (2019)
16. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016)
17. Zhao, D., Chen, Y., Lv, L.: Deep reinforcement learning with visual attention for vehicle classification. *IEEE Trans. Cogn. Dev. Syst.* **9**(4), 356–367 (2017)
18. Zhang, W., et al.: Coarse-to-fine UAV target tracking with deep reinforcement learning. *IEEE Trans. Autom. Sci. Eng.* **16**(4), 1522–1530 (2019)
19. Hajj, N., Awad, M.: On biologically inspired stochastic reinforcement deep learning: a case study on visual surveillance. *IEEE Access* **7**, 108431–108437 (2019)
20. Morrison, D., Corke, P., Leitner, J.: Learning robust, real-time, reactive robotic grasping. *Int. J. Robot. Res.* **39**(2–3), 183–201 (2019)
21. Wang, D.: SGDN: segmentation-based grasp detection network for unsymmetrical three-finger gripper. ArXiv, abs/2005.08222 (2020)
22. Wei, L., Joonwhaon, L.: A 3-D atrous convolution neural network for hyperspectral image denoising. *IEEE Trans. Geosci. Remote Sens.* **57**(8), 5701–5715 (2019)
23. Wang, Z., et al.: Accelerating atrous convolution with fetch-and-jump architecture for activation positioning. In: 2020 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), pp. 151–152. IEEE, Nanjing, China (2020)
24. Liu, Y., et al.: Adversarial learning for constrained image splicing detection and localization based on atrous convolution. *IEEE Trans. Inf. Forensics Secur.* **14**(10), 2551–2566 (2019)
25. Pan, X., et al.: An accurate nuclei segmentation algorithm in pathological image based on deep semantic network. *IEEE Access* **7**, 110674–110686 (2019)
26. Lv, L., et al.: Image semantic segmentation method based on atrous algorithm and convolution CRF. In: 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), pp. 160–165. IEEE, Dalian, China (2019)
27. Di Pan, et al.: Multi-oriented scene text detector with atrous convolution. In: 2020 Information Communication Technologies Conference (ICTC), pp. 346–350. IEEE, Nanjing, China (2020)

# Comparison of SAR Image Water Extraction Algorithms Based on Grey Incidence Analysis



Jingjue Chen , Rui Liu , Mei Yang , Xin Yang , Yuantao Yang , and Tianqiang Liu

**Abstract** The rapid extraction of water bodies is of critical significance to flood disaster assessment and water resource surveys. In recent years, the implementation of SAR images for water extraction has gradually attracted attention from scholars, but there are often pseudo water and “salt and pepper” phenomenon in the extraction results. In this study, the grey Level Co-occurrence Matrix (GLCM) was adopted to calculate 8 texture features of the SAR images, which increased the feature dimension of the model. Firstly, Grey incidence analysis (GIA) was employed for factor evaluation, and factors of higher importance were subsequently fed into Rotating forest (RF), k-Nearest Neighbor (KNN), and Logistic Regression (LR) models to build DGI based hybrid models named DGI-RF, DGI-KNN and DGI-LR models respectively. Finally, the models were used to conduct a large-scale water extraction in Pengze County, China to evaluate the generalization ability of the Models. Results based on fivefold verification show that DGI-RF model has the best generalization ability, followed by DGI-KNN model, and DGI-LR model has worst generalization ability. It is worth noting that RF and KNN coupled with DGI can effectively alleviate the interference of mountain shadows, speckle noise, and salt and pepper phenomenon, which may be computationally efficient and accurate alternatives for large-scale water extraction.

**Keywords** SAR image · The Grey level co-occurrence matrix · Grey incidence analysis · Water extraction

---

J. Chen

College Earth Sciences, Chengdu University of Technology, Chengdu 610059, China

R. Liu · M. Yang · X. Yang · Y. Yang · T. Liu

College of Geophysics, Chengdu University of Technology, Chengdu 610059, China

e-mail: [lr@cdut.edu.cn](mailto:lr@cdut.edu.cn)

## 1 Introduction

Water extraction technology based on remote sensing images can quickly obtain water distribution information in dangerous areas [1], which has critical significance for land and resources management, water resources investigation, drought and flood disaster prevention, and rapid flood disaster assessment [2]. However, optical remote sensing satellite monitoring is feasible only when there is a light source, and the received wavelength range is relatively short. The SAR remote sensing satellite adopts an active imaging method, which can receive long waves with strong penetrating power without a light source, so it can observe the earth all-time and all-weather [2, 3]. The long waves emitted by SAR satellites can penetrate clouds and fog to obtain water information under extreme conditions, while the water in the optical images generally presents different spectral characteristics, which inevitably increases the difficulty of mapping the water distribution [4]. Due to the insensitivity of SAR to clouds, rain and fog and the reflection of water bodies were mainly specular reflections, SAR images in areas with heavy rainfall were obtained in this study and research on water body information extraction was carried out.

Currently, there are two main categories of water body extraction methods based on SAR images: threshold segmentation method and classification method [5]. Among them, the most representative threshold segmentation methods are OTSU algorithm and entropy threshold method [1]. Many research [6] have implemented the OTSU to extract water bodies, but there were usually many pseudo-water bodies. These pseudo-water bodies are generally composed of mountain shadows, roads and speckle noise. In order to solve the problem of mis-segmentation caused by these factors, post-processing methods based on morphology or DEM are usually used to remove pseudo water pixels to obtain a more accurate water distribution map [4]. Classification mainly refers to the extraction of water information based on machine learning algorithms and object-oriented classification methods [5]. Scholar [7] selected two GF-1 satellite images of different scales and complexity in the Poyang Lake area, and compared the extraction effects of SVM, object-oriented method and water index method in the study area, indicating that SVM has the highest extraction accuracy. Due to the fact that pixel-based classification is prone to “salt and pepper phenomenon” [8], scholar [9] used the texture features calculated by the grey-level co-occurrence matrix to establish a multi-dimensional feature space, which effectively reduced this phenomenon. The ensemble learning classification method can effectively reduce the uncertainty in the classification process by integrating multiple classifiers, thereby improving the accuracy of image classification [10]. The rotating forest algorithm proposed by Rodriguez in 2006 is a new type of ensemble learning algorithm [11], which has been widely used in data classification since it was proposed. It has also been verified that the rotating forest model is superior to the random forest (RF) algorithm and AdaBoost [12]. Among various classic machine learning models, KNN is a classifier based on distance iterative analysis, while the LR model is a linear classification model. These two models have achieved satisfactory performance when applied to remote sensing image classification [13, 14].

In previous studies, mathematical statistical methods such as regression analysis, analysis of variance, and principal component analysis are generally used to analyze the relationship between various factors in the system. Although statistical methods can solve many practical problems, they usually require a large sample size and require the data to have a typical probability distribution, which is often difficult to meet in practical applications. The grey incidence analysis (GIA) method proposed by Professor Deng Julong (1985) is widely used in economics, medical education, and geology [15]. GIA is an important part of grey system theory. Its basic idea is to judge whether the connection between different sequences is close according to the geometric shape of the sequence curve. With fewer samples, the degree of grey incidence (DGI), which indicates the importance of the comparison sequence and the reference sequence, can be analyzed.

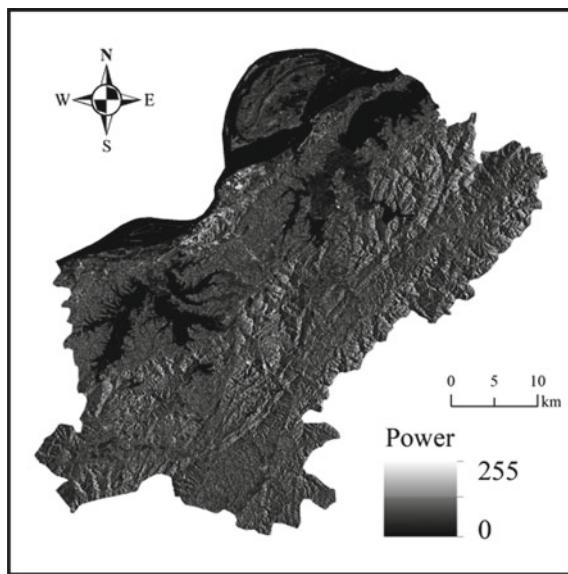
The primary objective of this paper is to add texture features to alleviate the influence of mountain shadows, speckle noise, roads and other factors on water extraction, and effectively reduce shadows and "salt and pepper phenomenon" in the extraction results. When extracting texture features, there are numerous factors need to be considered, but few studies have explored the impact of number of factors on water body extraction. Therefore, based on the factors obtained from the GIA evaluation, the GIA-RF, GIA-KNN and GIA-LR models were constructed to compare the water extraction performance, and finally the optimal model was obtained.

## 2 Materials

### 2.1 Study Area and SAR Data

The study area in this paper is located in Pengze County, Jiangxi Province, China, which is between  $29^{\circ}34'32''$  N ~  $30^{\circ}04'50''$  N,  $116^{\circ}21'58''$  E ~  $116^{\circ}53'48''$  E. The area is about  $1542 \text{ km}^2$ , and the terrain is high in the south and low in the north. The southeastern part of this area is mountainous, the middle part is hilly, and the northwest along the Yangtze River is an alluvial plain. The SAR image of this area is shown in Fig. 1.

The SAR image used in our study is retrieved from the free dataset of Sentinel-1A satellite of ESA with a spatial resolution of 25 m. The overpass time of the SAR image is July 20, 2020. Then a Series of operations such as orbit refining, multilooking, filtering, radiation correction, and geocoding were performed. All data used in this study are unified to the WGS-84 datum.

**Fig. 1** Power image

## 2.2 Texture Factors and DEM

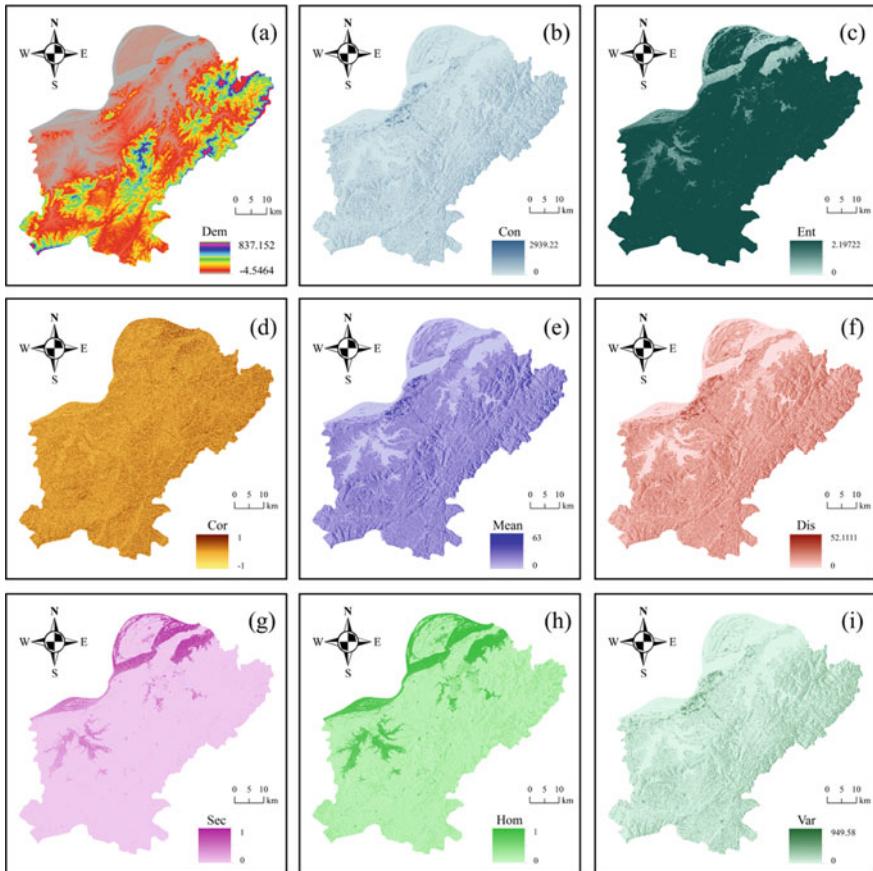
The grey Level Co-occurrence Matrix (GLCM) describes the texture characteristics of the image by mining the spatial correlation characteristics of the grey level of the image. GLCM is usually calculated in four directions of  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$  [16]. In this study, a window size of  $3 \times 3$  was set and 8 texture features were calculated, namely Contrast (Con), Correlation(Cor), Dissimilarity (Dis), Entropy (Ent), Homogeneity (Hom), Mean, Second Moment (Sec), Variance (Var). The texture factors are shown in Fig. 2a–i.

The 90 m resolution SRTM DEM data obtained in this study was resampled to 25 m by the bicubic convolution method, in order to be consistent with the preprocessed SAR image resolution (25 m).

## 3 Methodology

### 3.1 Grey Incidence Analysis Theory

The basic idea of GIA is to transform the discrete behavior observations of system factors into piecewise continuous polylines by linear interpolation, so as to construct a model to measure the correlation degree based on the geometric characteristics of the polylines.



**Fig. 2** Map of Factor

- (1) The dimensionlessness of the input sequence: the dimensionlessness is the prerequisite for the grey incidence analysis. The factors data are processed using the normalization method. The Normalization operation can be defined as formula (1):

$$x_i(k) = \frac{x_i(k) - \min(x_i(k))}{\max(x_i(k)) - \min(x_i(k))} \quad (1)$$

where the range of factor values is calculated to [0,1]. The data in this study are all normalized, so GIA calculations can be performed directly.

- (2) Reference sequence and comparison sequence. Assume that the water label is the reference sequence  $X_0 = \{x_0(k), k = 1, 2, \dots, n\}$ , which  $x_0(k) = 1$ . Assume that the comparison sequence  $X_i = \{x_i(k), k = 1, 2, \dots, n\}$  consists of factors.

- (3) Grey incidence coefficient (GIC): the GIC is the premise for calculating DGI, which can be defined as:

$$v_{\min}(i, k) = \min_i \min_k |x_0(k) - x_i(k)| \quad (2)$$

$$v_{\max}(i, k) = \max_i \max_k |x_0(k) - x_i(k)| \quad (3)$$

$$r(x_0(k), x_i(k)) = \frac{v_{\min}(i, k) + \xi v_{\max}(i, k)}{|x_0(k) - x_i(k)| + \xi v_{\max}(i, k)} \quad (4)$$

where  $r(x_0(k), x_i(k))$  A is the incidence coefficient at point  $k$ ;  $\xi$  is the distinguishing coefficient, the value which is generally 0.5.

- (4) Degree of grey incidence represents the correlation degree between each subsequence and the parent sequence, and a large grey correlation degree indicates higher factor importance. It can be expressed as:

$$r(X_0, X_i) = \frac{1}{n} \sum_{k=1}^n r(x_0(k), x_k(k)) \quad (5)$$

where the value of  $r(X_0, X_i)$  is between [0,1], the importance degree of the factors can be obtained by ranking the degree of grey incidence.

### 3.2 Water Extraction Model

Based on the DIA ranking, a combination of factors was selected as the feature space for constructing a water body extraction model. As for the water body extraction model, three different types of machine learning models were employed, namely the RF algorithm based on decision trees [12], the distance-based KNN algorithm [13] and the LR based on linear regression [14].

## 4 Results

### 4.1 Evaluation of Factors

The DGI between the relevant factors and the water body is calculated by GIA, and the calculation results are shown in Tables 1 and 2. Subsequently, the DGI is sorted in descending order, and the results are as follows:

$$\text{Con} > \text{Var} > \text{Mean} > \text{Dis} > \text{Power} > \text{Dem} > \text{Ent} > \text{Sec} > \text{Cor} > \text{Hom} \quad (6)$$

**Table 1** Results of DGI

Factor	Con	Cor	Dem	Dis	Ent
DGI	0.9979	0.5009	0.9778	0.9845	0.6094

**Table 2** Results of DGI

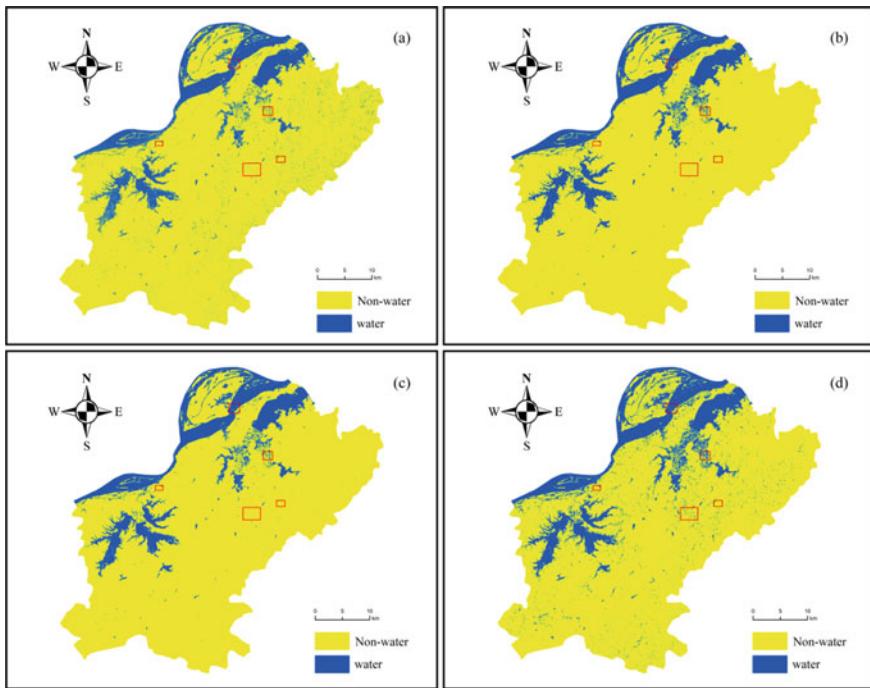
Factor	Hom	Mean	Power	Sec	Var
DGI	0.3803	0.9910	0.9795	0.5020	0.9978

From the importance ranking result and the DGI value in Tables 1 and 2, it can be concluded that the DGI value is mainly concentrated in 0.9 ~ 1.0 and 0.3 ~ 0.7. In the expression (6), the DGI values of the factors Con ~ Dem are all greater than 0.9, while the DGI values of the four factors Ent ~ Hom are all less than 0.7. In GIA, the factors with DGI values between 0.7 and 1.0 are considered to be the most significant factors, and DGI values less than 0.7 are considered to be less significant ones. Finally, six factors with relative high importance (Con, Var, Mean, Dis, Power, and Dem) were selected to participate in the construction of the model, and the DGI of each factor is greater than 0.9.

#### 4.2 Results of Water Extraction

The combination of factors selected by the GIA were then fed into the RF model, KNN model and LR model for training. The three models trained with these six factors were called DGI-RF, DGI-KNN and DGI-LR. Simultaneously, the Power map was adopted to construct a single-factor RF model as the Power-RF model.

In order to ensure the consistency of the training sample area, the training sample area of the Power-RF model is consistent with the training sample area of the DGI-RF, DGI-KNN and DGI-LR models. The extraction results of the four model water bodies are shown in Fig. 3. According to the results in the figure, it can be concluded that DGI-RF and DGI-KNN are better than Power-RF and DGI-LR in handling mountain shadows. Although the DGI-LR model is weaker than DGI-RF and DGI-KNN in processing hill shades, it has certain processing capabilities in some areas. In order to facilitate the discussion, five enlarged rectangular areas corresponding to different situations are selected as shown in Fig. 3.



**Fig. 3** Figure of water extraction results; **a** Map of Power-RF, **b** Map of DGI-RF, **c** Map of DGI-KNN, **d** Map of DGI-LR

#### 4.3 Model Performance and Accuracy Evaluation

In this study, the fivefold was employed to evaluate the performance of each model. The fivefold average overall accuracy and 95% confidence interval of the four models are shown in Table 3.

It can be seen from Table 3 that the 95% confidence interval is four digits after the decimal point, which proves the stable generalization ability of the model. The model with the highest accuracy is applied to the test set, and the performance of the four models on the test set is shown in Table 4. The sorting results of various evaluation indicators are shown in Expressions (7) to (10).

**Table 3** Result of K-fold Cross-validate

Method	Validate OA
Power-RF	0.9637 ( $\pm 0.0032$ )
DGI-RF	0.9996 ( $\pm 0.0004$ )
DGI-KNN	0.9982 ( $\pm 0.0007$ )
DGI-LR	0.9752 ( $\pm 0.0025$ )

**Table 4** Evaluation of the test set accuracy

Method	Precision	Recall	F1	Kappa
Power-RF	0.9825	0.9600	0.9711	0.9208
DGI-RF	0.9978	0.9997	0.9987	0.9965
DGI-KNN	0.9979	0.9993	0.9986	0.9961
DGI-LR	0.9976	0.9617	0.9793	0.9440

Among them, the accuracy rankings of the four models in Recall, F1, and Kappa are consistent. In Precision, the accuracy of DGI-KNN is higher than that of DGI-RF, but from other indicators, DGI-RF is higher than all models. Judging from the performance on the test set, the indicators of the four models are all higher than 0.9, indicating that the models all have good predictive ability and can be used for large-scale water extraction.

$$\text{Precision: DGI - KNN} > \text{DGI - RF} > \text{DGI - LR} > \text{Power - RF} = 0.9825 \quad (7)$$

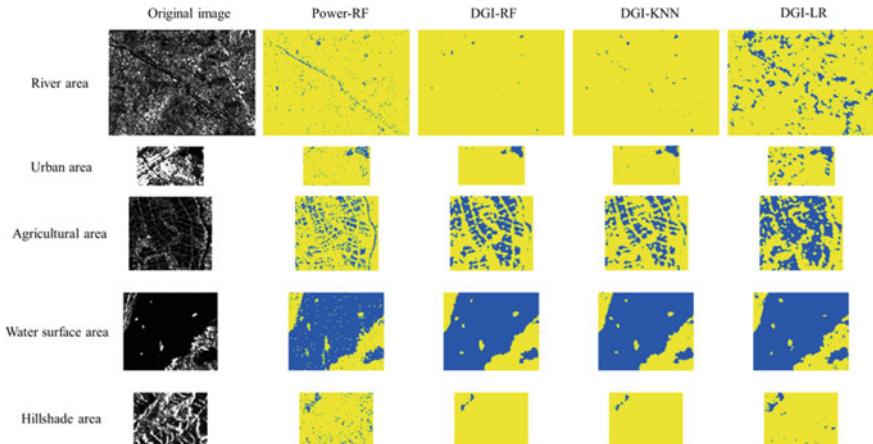
$$\text{Recall: DGI - RF} > \text{DGI - KNN} > \text{DGI - LR} > \text{Power - RF} = 0.9600 \quad (8)$$

$$\text{F1 : DGI - RF} > \text{DGI - KNN} > \text{DGI - LR} > \text{Power - RF} = 0.9711 \quad (9)$$

$$\text{Kappa: DGI - RF} > \text{DGI - KNN} > \text{DGI - LR} > \text{Power - RF} = 0.9208 \quad (10)$$

## 5 Discussion

Figure 4 shows an enlarged view of the five rectangular areas in Fig. 3, corresponding to the river, city, farmland, water surface, and mountain shaded areas. It can be seen that Power-RF can correctly distinguish the outline of the small river in the River area, while the three models constructed based on DGI cannot distinguish well. The DGI-LR model even has many mis-extractions in this area. In urban areas, DGI-RF and DGI-KNN can accurately distinguish roads, so that the extraction results are hardly affected by roads, while Power-RF and DGI-LR have a weaker ability to distinguish roads. In agricultural areas, DGI-RF and DGI-KNN can accurately identify the contours of submerged farmland, while DGI-LR has mis-extraction, and Power-RF has salt and pepper phenomenon. In the wide surface area of the water body, the four models all distinguished the ships on the water surface successfully. Compared with Power-RF, the salt-and-pepper phenomenon did not appear in the three models, which shows that the hybrid models based on GIA can alleviate this phenomenon. In the mountain shadows area, DGI-RF and DGI-KNN can effectively



**Fig. 4** Enlarge the display

remove the influence of the mountain shadows, while the other two models failed to achieve the desired results.

The Power-RF model did not distinguish the mountain shadows accurately. In detail, there are many small spots in the extraction results, but for small rivers, Power-RF has a certain ability to distinguish. The possible reason is that when single image is used for modeling, the image characteristics of the water body are very similar to the image characteristics of the shadow of the mountain shadows. In addition, the captured image environment is in a period of heavy rain, odd scattering will occur on the ground surface, resulting in mountain shadows showing similar features to water bodies in the image. Light winds and strong winds may cause changes in the reflection coefficient of the water body, which may also cause stray points to be distributed on the water surface extracted by the Power-RF model established using a single image.

The three DGI-based models were all based on the factors selected by the GIA, but the generalization ability of the DGI-LR model was poor in farmland areas. The possible reason is that the LR model is a linear model, and the field stalks around the farmland may also have odd scattering. Since the texture features of factors were established based on the original image, the mountain shadows are similar to the feature of the water body. Therefore, it was extremely possible that similar texture features had appeared, which will weaken the generalization ability of the DGI-LR model. Secondly, the factors and reference sequences obtained by GIA will appear non-linear, which may also be another inducement for more mis-extraction in paddy fields and mountainous areas. Moreover, the performance of DGI-LR on the test set was only slightly better than that of Power-RF, which shows that the generalization ability of DGI-LR is similar to that of Power-RF, as shown in Fig. 3a and d.

When extracting texture features, the information of surrounding pixels needs to be considered, but the width of a small river is usually only 2 to 3 pixels. After

processing, the information of small rivers may have been lost, resulting in a poor prediction ability for small rivers based on DGI models.

## 6 Conclusion

The single-factor model constructed using single Power image is insensitive to mountain shadows and speckle noise. In this paper, the GIA method is adopted to calculate the DGI of the reference sequence and the comparison sequence, with 0.7 as the importance threshold. Among those hybrid machine learning models, DGI-RF and DGI-KNN can effectively alleviate the influence of mountain shadows, speckle noise, and environmental changes. However, the two proposed models cannot distinguish small rivers of only 2–3 pixels. In conclusion, the DGI-RF model with texture features is promising in large-scale water extraction which can provide a reference idea for water extraction based on SAR images.

## References

1. Dan, L., Baosheng, W., Bowei, C., et al.: Review of water body information extraction based on satellite remote sensing. *J. Tsinghua Univ. (Sci. Technol.)* **60**(02), 147–161 (2020)
2. Qi, Z., Yuanbo, L., Jing, Y., et al.: Lake hydrology in China: advances and prospects. *J. Lake Sci.* **32**(05), 1360–1379 (2020)
3. Rui, H., Shaoping, D., Liya, Z.: Experiment and method of TerraSAR orthorectification based on precise orbital data. *Sci. Surv. Mapping* **40**(10), 153–156 (2015)
4. Xinzhi, G., Qingwei, Z., Hua, S., et al.: Study on water information extraction using domestic GF-3 image. *J. Remote Sens.* **23**(03), 555–565 (2019)
5. Longfei, S., Zhengxuan, L., Fei, G., et al.: A review of remote sensing image water extraction. *Remote Sens. Land Resour.* **33**(01), 9–19 (2021)
6. Yu, L., Yun, Y., Quanhua, Z.: Waterbody extraction from SAR imagery based on improved speckle reducing anisotropic diffusion and maximum between-cluster variance. *J. Geo-inf. Sci.* **21**(6), 907–917 (2019)
7. Qiuya, D., Lingkui, M., Zhiwei, F., et al.: Applicability of the water information extraction method based on GF-1 image. *Remote Sens. Land Resour.* **27**(04), 79–84 (2015)
8. Peng, Z., Yuanli, X., Guangxin, J., et al.: Advances on water body information extraction from remote sensing imagery. *Remote Sens. Inf.* **35**(05), 9–18 (2020)
9. Jing, L., Yunhao, C., Weiguo, J.: Water and settlement area extraction from single-band, single-polarization SAR image based on SVM method. *J. Image Graph.* **02**, 257–263 (2008)
10. Pejun, D., Alim, S.: Multiple instance ensemble learning method for high-resolution remote sensing image classification. *J. Remote Sens.* **17**(01), 77–97 (2013)
11. Junshi, X.: Hyperspectral Remote Sensing Image Classification Based on Ensemble Learning. China University of Mining and Technology (2013)
12. Rodriguez, J., Kuncheva, L., et al.: Rotation forest: a new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* (2006)
13. Bing, T., Xiaofei, Z., Guoyun, Z., et al.: Hyperspectral image classification via recursive filtering and KNN. *Remote Sens. Land Resour.* **31**(01), 22–32 (2019)
14. Bué, I., Catalo, J., Semedo, L.: Intertidal bathymetry extraction with multispectral images: a logistic regression approach. *Remote Sensing* **12**(8), 1311 (2020)

15. Wenjie, D., Sifeng, L., Zhigeng, F.: On modeling mechanisms and applicable ranges of grey incidence analysis models. *Grey Syst. Theory Appl.* **8**(04), 448–461 (2018)
16. Haralick, R.M.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern. SMC* **3** (1973)

# Combine Local and Global Feature Extraction for Point Cloud Classification



Xiaolong Lu<sup>ID</sup>, Baodi Liu<sup>ID</sup>, Weifeng Liu<sup>ID</sup>, Kai Zhang<sup>ID</sup>, Ye Li<sup>ID</sup>,  
and Peng Liu<sup>ID</sup>

**Abstract** The point cloud is one of the common formats of 3D data, and it can represent the shape of objects more intuitively. However, due to the point clouds' irregularity and disorder, there are still many problems during processing. The previous approach was to convert point clouds to other formats for processing until PointNet came along, which pushed point cloud data directly into network processing for the first time, achieving a breakthrough. The conventional approaches to dealing with point clouds rarely simultaneously consider the global and the local features of point clouds. With the appearance of attention mechanism and graph structure, the application on point cloud also has a certain effect. In particular, the graph structure is more suitable for the processing of the point cloud due to its characteristics. In this paper, the attention mechanism is the basis to enhance the representation of nodes, and then the dynamic graph and point network are fused to extract local and global features, respectively. Finally, we conducted experimental verification on the benchmark datasets, such as ModelNet40 and ScanObjectNN, and achieve superior performance to several state-of-the-art approaches.

**Keywords** Point cloud classification · Attention mechanism · Dynamic graph · Local and global feature

---

X. Lu

College of Oceanography and Space Informatics, China University of Petroleum (East China),  
Qingdao, China

B. Liu (✉) · W. Liu

College of Control Science and Engineering, China University of Petroleum (East China),  
Qingdao, China

e-mail: [liubaodi@upc.edu.cn](mailto:liubaodi@upc.edu.cn)

P. Liu

Shandong Kexun Information Technology Co., Ltd., Qingdao, China

Y. Li

Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

K. Zhang

School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China

## 1 Introduction

With the increasing application of 3D data, a series of sensors, such as lidar sensors, emerge at the historical moment, which acquires 3D data more efficiently and effectively. As the most representative point cloud data in 3D data, because it can best reflect the original sensor data, it has become more and more important, and its use has also increased. It has become particularly popular in areas such as autonomous driving, robot recognition, and unmanned aerial vehicles. However, due to the irregularity, disorder, and sparsity of point cloud data, there are certain challenges in actual processing.

The traditional work is to transform point cloud data into other data formats: Based on multi-view and volumetric grid.

For Multi-view, the MVCNN [2] is a pioneering work based on multi-view. Its innovation is to use 2D rendering images obtained from 3D data of objects from different “perspectives” as original training data and then apply a 2D convolution operation training model to achieve a better classification effect. MHBN [3] is used to coordinate bilinear pools to aggregate local convolution features to obtain an effective representation of 3D objects. For the volumetric grid, VoxNet [1] uses 3D CNN to process the voxel of the occupied grid, which can2 quickly and accurately classify 3D data. In order to make a 3D convolutional neural network effective in existing models, OctNet [4] divides sparse data into a series of hybrid octrees by layers, which enables the network to achieve both deep level and high resolution. PointGrid [5], which is a three-dimensional convolutional network, is the integration of points and grids and is a hybrid model that can better represent the details of local geometry. But in the process of conversion, important data information is easily lost, and unnecessary data redundancy is introduced, which increases the computational burden.

With the development of scientific research technology, PointNet [6] emerged, making direct use of point cloud data. This network uses multi-layer perceptron and symmetric functions to ensure the disorder and permutation invariance of the point cloud. Since the network is trained to point by point, it does not consider the local feature extraction. Later PointNet++ [7] is optimized based on PointNet, and the most distant point sampling method is adopted to realize hierarchical feature extraction so as to obtain fine-grained features from the neighborhood of each point. But PointNet++ only uses local ball query to do max pooling, and there is no deeper exploration of the local feature information of the point cloud.

In order to solve the problem that the information between points cannot be considered, the graph structure is introduced into point cloud processing. DGCNN [8] first uses k-NN to construct the graph structure between points, analyzes the geometric relations between points more accurately, and then uses the edge convolution module to fully extract the local information. Further, LDGCNN [17] removes the transformation network in DGCNN and is inspired by Densenet to connect different levels of features and improve the network’s performance. Zhang et al. [9] proposed a new network based on graph convolution, which mainly used the graph convolution form

of Chebyshev polynomial to process the point cloud features and finally realized the classification through the full connection layer. Due to the uniqueness of the point cloud, the effectiveness of attention mechanisms in various fields, and the influence of graph attention mechanism, GAPNet [10] introduced graph attention mechanism into the point cloud processing task. However, in the process of utilization, only point-based network feature extraction was used in the later processing, which could not extract more fine-grained features.

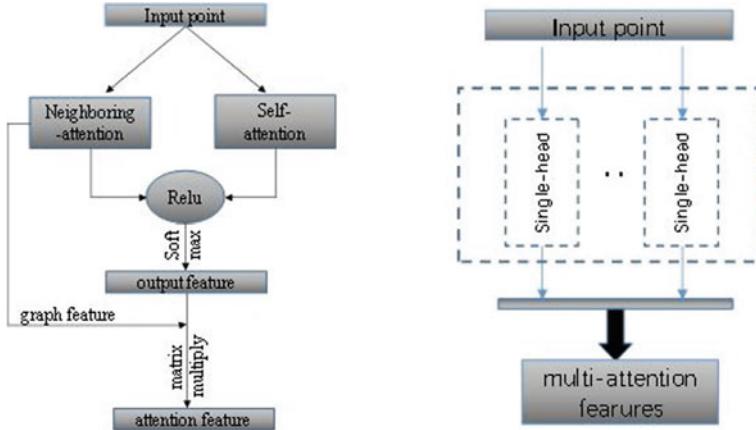
In this paper, to overcome the inadequacies of point cloud feature extraction, we extract the attention feature based on self-attention and neighboring-attention and then extract the fine-grained feature of the point cloud from local and global aspects so that the context information of point cloud and the global information of point cloud can be fully considered. The 3D coordinates of the point cloud are taken as input and then through the transformation network with attention. On the one hand, the dynamic graph network is used to further process the point cloud, which enables the network to extract more detailed local neighborhood information. On the other hand, let the point cloud go through a multi-head parallel attention mechanism and then through a stacked multi-layer perceptron (MLP) with Shared parameters to extract global features. Finally, a graphic-based pooling layer is connected to further enhance the robustness of the network.

The main contributions of this paper are as follows:

- (1) It is easy to ignore the structural information between dropped points when only considering the feature extraction of points. In this paper, global and local features are considered at the same time so that more fine-grained information can be mined.
- (2) In this paper, on the basis of including the attention mechanism, we combine the dynamic graph structure with the Shared perception machine module with jump connection to get a better effect.
- (3) We tested our network on the classified data sets ModelNet40 and ScanObjectNN and showed the performance of the network in the experimental part.

## 2 Methodology

In this section, we describe the architecture that our network uses to better handle irregular data, such as point clouds, which are mainly used for point cloud classification processing tasks. We explained the method used in this paper in detail, mainly including the following three parts: attention mechanism module, dynamic graph structure, point network structure, and the network model we designed is shown in Fig. 1.



**Fig. 1** GAPLayer. For the left part, the single-headed GAPLayer first extracts the self-attention and neighboring-attention features of the point cloud, respectively, and then performs a normalization operation through a nonlinear activation function LeakyRelu and softmax functions, and finally gets the final attention characteristics by multiplying the graph features. The right part represents a multi-headed attention mechanism, and 4-heads are used in this paper

## 2.1 Attention Mechanism Module

The attention machine used in this paper considers the local and global structure and uses the self-attention mechanism and the neighboring-attention mechanism, as shown in the left part of Fig. 1. Through the self-attention mechanism, the importance of independent points is better analyzed, and corresponding weight information is given, which is represented by Eq. 1:

$$x'_i = g(x_i, \theta) \quad (1)$$

where,  $x'_i$  represents the output features encoded on the point features,  $g()$  represents a nonlinear function (we use a multi-layer perceptron with shared parameters), while  $\theta$  represents a learnable filter parameter. For the neighbor attention module, we first use the k-NN (k-nearest neighbor algorithm) to construct a graph structure  $G = (V, E)$  to represent the relationship between points, where  $V = (1, 2, \dots, N)$  is the number of points of a point cloud object,  $E$  is the edge information connecting adjacent points. We defined the feature information of this part as  $y_{ij} = (x_i - x_{ij})$ , where  $i$  and  $j$  respectively represent node and neighbor point indexes,  $x_{ij}$  is the neighboring point  $x_j$  to point  $x_i$ . Similar to Eq. 1, this part can be expressed by the following formula:

$$y'_{ij} = g(y_{ij}, \theta) \quad (2)$$

Through the above two formulas, we can obtain the self-attention feature  $x'_i$  and the neighboring-attention feature  $y'_{ij}$ , the attentional features can then be obtained

from Formula 3, where LeakyReLU() represents a nonlinear activation function.

$$c_{ij} = \text{LeakyReLU}\left(g(x'_i, \theta) + g(y'_{ij}, \theta)\right) \quad (3)$$

To better explore the attention features of the point cloud and more local and global information, we also constructed a multi-head attention structure (Fig. 1 (right)) based on the above attention module and ran the multi-head attention module in parallel.

## 2.2 Feature Extraction from Point Network

In this section, we first apply the multi-head attention mechanism mentioned in the previous section to obtain features with attention coefficients. Inspired by PointNet [6], the feature with the attention coefficient is input into an MLP with shared parameters to extract the fine-grained feature at the point. In the process of its implementation, ResNet [11] is used for reference, and a skip connection is adopted, which makes the input contain more layers of information. To improve the network's performance, we also introduce an attention-pooling operation to identify the most important part of the multi-head attention feature.

## 2.3 Feature Extraction from the Graph Structure

We know that point-based feature extraction cannot capture the context information of the point cloud well, nor can it achieve a better classification effect. Inspired by DGCNN [8], we fuse the dynamic graph structure better to connect the context information of the point cloud and facilitate the extraction of fine-grained features.

In this part, we first construct a directed graph  $G = (V, E)$  to represent the local structure between point clouds, where  $V = (1, 2, \dots, N)$  and  $E \subseteq V \times V$  indicate vertex and edge information about a point cloud, respectively. At the same time, the edge feature is defined as  $ase_{ij} = g_\Theta(x_i, x_j)$ , where  $g : R^C \rightarrow R^{C'}$  represents a nonlinear function with a set of trainable parameters. Finally, a convolution operation is defined on the graph, and an asymmetric channel-based aggregation function is used (for example,  $\Sigma$  or pooling). Simply put, input a 3D point cloud with  $n$  points, and this operation generates a  $nC'$ -dimensional point cloud with  $n$  points. Therefore, the graph convolution output of the  $i$ -th node can be expressed by the following formula:

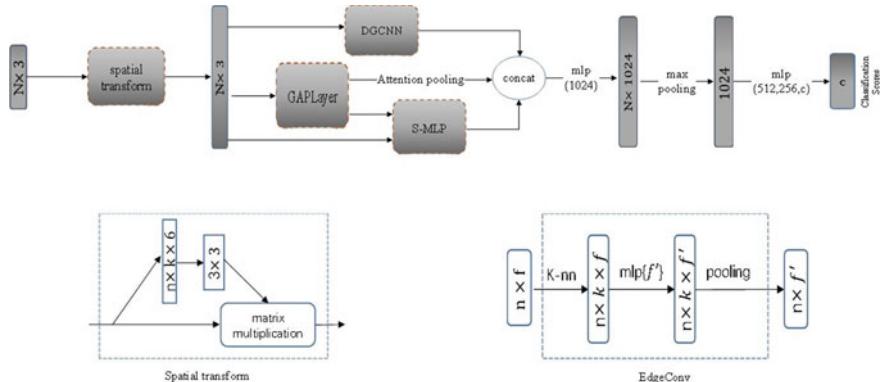
$$x'_i = \Delta_{j:(i,j) \in E} g_\Theta(x_i, x_j) \quad (4)$$

The choice of the nonlinear function  $g$  and the aggregate function  $\Delta$  is also crucial. In the previous approach, some researchers tried to encode only the global information and ignore the local structure information, such as the method in PointNet. Later, some people said that the information was divided into small pieces, which fully considered the local information but lost the global information. In this work, we calculate the global structure information and local structure information through the coordinate information of and  $x_j - x_i$  respectively, and the output of one layer can be expressed by the following formula:

$$g_\Theta(x_i, x_j) = \bar{g}_\Theta(x_i, x_j - x_i) \quad (5)$$

## 2.4 Our Network

Inspired by GAPNet [10] and DGCNN [8], we proposed the model of this paper, as shown in Fig. 2. In this model, we first use a 4-head attention module to deal with the point cloud so that the output of the transformation network has certain attention



**Fig. 2** *The network model:* The proposed network model is mainly used for classification tasks. The classification model takes  $N$  as input and only considers the 3D coordinate ( $x, y, z$ ) of the point cloud. The upper part uses a dynamic graph module to extract the local information of the point cloud. And the lower part first uses a GAPLayer to obtain an attention feature and graph feature. Here, multi-graph features are followed by an attention pooling operation, and multi-attention features are followed by a perceptron with shared parameters to extract point-based features. Finally, global features, local features, and attention pooling features are combined to obtain the classification score of category  $c$ . *Spatial transformation network:* The use of this network guarantees the permutation invariance of the point cloud. The model first uses a single-head attention module and finally generates a  $3 \times 3$  transformation matrix. *Edge convolution:* This module is mainly a structure used in dynamic graphs (DGCNN). First, a graph structure is constructed using k-NN. Then edge feature information is calculated through the perceptron module with shared parameters and pooling operations

features and keeps some transformation invariance to the point cloud. Next, a multi-head attention module is used to extract feature attention information and apply it to a multi-layer perceptron with a jump connection structure to extract more fine-grained features. At the same time, we also introduce a dynamic graph structure and run it in parallel with the above parts to extract local information between point clouds, thus ensuring that our model carefully considers both local and global information. Finally, an attentional pooling operation is introduced to make the entire network model more robust and improve performance.

### 3 Experiments

In this part, we first introduce two point cloud datasets commonly used for classification and verify our model on these two datasets. Then our method is compared with existing methods in the CAD model and real-world point cloud dataset. Finally, we introduce the ablation experiment to prove the strong performance of our network.

**Datasets.** We experimented with two datasets and demonstrated the performance of our network: a synthetic CAD model classification dataset ModelNet40 [19] and a real-world point cloud dataset ScanObjectNN [20].

**ModelNet40.** As a classical point cloud classification data set, its application in recent years has also received great attention. The ModelNet40 data set contains 12,311 synthetic grid CAD models, which are composed of 40 categories. Among them, the training data contains 9843 models, and the remaining models are used for experimental testing. In addition, 1024 points were uniformly sampled on the surface of each model, and the model was further normalized through the unit sphere. For simplicity, the 3D coordinates (x, y, z) of the point cloud were only used in the experiment of this paper.

**ScanObjectNN.** Experiments on synthetic datasets have achieved relatively high performance; however, experiments on real-world point cloud datasets remain a challenge. In this paper, we have done experiments on the ScanObjectNN data set and made a comparison. The data set has about 15,000 objects, divided into 15 categories of information. Since the data set reflects the point cloud data in the real world, there are some missing parts and the influence of the background, which makes the data set according to the challenge.

**Training.** We use adam optimization model with a momentum of 0.9, and the learning rate is set to 0.001, and the batch size is 32 and trained 250 epochs. The decay rate of batch normalization was initially 0.7 and gradually increased to 0.99 during training. Our model is trained on Nvidia Tesla V100 GPU and TensorFlow-GPU 1.14.

**Experimental results on ModelNet40.** Experimental results on the dataset are shown in Table 1. This table shows several recently popular point cloud classification methods. It is obvious that our method has better experimental results than

**Table 1** Experimental results on ModelNet40

Method	Avg class acc. (%)	Overall acc. (%)
ECC [12]	83.2	87.4
PointNet [6]	86.0	89.2
PointNet+ + [7]	87.8	90.7
SO-Net [13]	87.3	90.9
KCNet [14]	–	91.0
3DmFV [15]	86.3	91.4
DGCNN [8]	90.2	92.2
GAPNet [10]	89.7	92.4
<b>Ours</b>	<b>90.3</b>	<b>92.7</b>

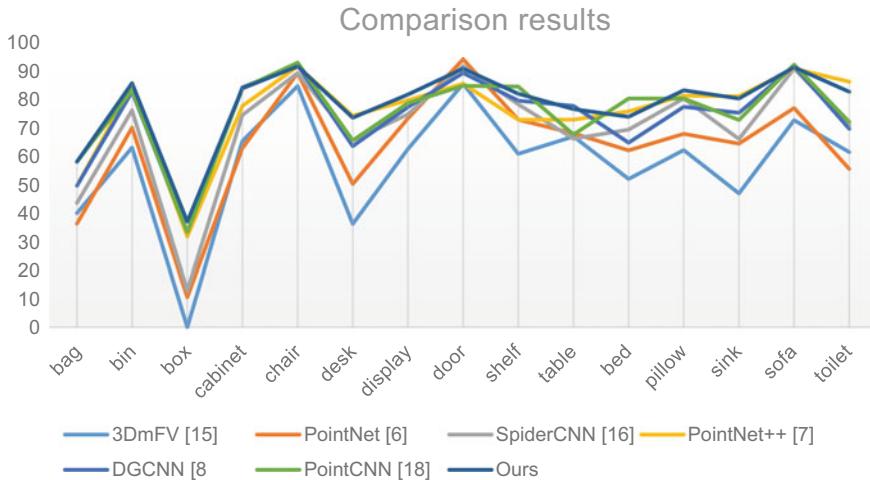
the compared methods, which is 0.5% higher than DGCNN and 0.3% higher than GAPNet. The symbol ‘–’ means the results are unavailable.

**Experimental results on ScanObjectNN.** With the same experimental setup, we also verified our model on ScanObjectNN, the real-world dataset, as shown in Table 2. As can be seen from the table, the overall accuracy of our experiment has reached 80.7%, which is significantly better than other methods. This experiment fully demonstrates that the effectiveness and robustness of our model in the face of a real-world dataset are exactly what we need now. In Fig. 3, we show the accuracy of each category and can clearly see that our proposed method is better than the other methods compared through experiments.

**Ablation studies.** In order to verify the validity of our model, we also carried out ablation experiments on ModelNet40 and ScanObjectNN, respectively, as shown in Table 3. On ModelNet40, we compare the proposed method with that in garnet. It can be seen from the table that the accuracy of the proposed method is reduced by 0.2% when the original attention pooling layer is replaced by a dynamic graph. When we run dynamic graphs in parallel, our experimental results are improved by 0.3%. Ablation studies on ScanObjectNN are shown in the third column of the table.

**Table 2** Experimental results on ScanObjectNN

Method	Avg class acc. (%)	Overall acc. (%)
3DmFV [15]	58.1	63
PointNet [6]	63.4	68.2
SpiderCNN [16]	69.8	73.7
PointNet+ + [7]	75.4	77.9
DGCNN [8]	73.6	78.1
PointCNN [18]	75.1	78.5
<b>Ours</b>	<b>77.8</b>	<b>80.7</b>



**Fig. 3** Comparison of the accuracy of each class on the dataset ScanObjectNN

**Table 3** Ablation experiments

Components	ModelNet40	ScanObjectNN
GAPlayer + attention pooling	92.4	76.2
GAPlayer + DGCNN	92.2	80.4
GAPlayer + DGCNN + attention pooling	<b>92.7</b>	<b>80.7</b>

## 4 Conclusion

In this paper, we propose a new model for the point cloud classification task. We fully consider the local and global information of the point cloud and extract the information through the dynamic graph structure and the point network structure with attention features, respectively. Moreover, the method in this paper is still based on the graph attention mechanism module, and the extracted features with attention coefficient are further processed. Finally, an attention pooling operation is introduced to ensure the effectiveness and robustness of the network. Our model is validated on ModelNet40 and ScanObjectNN datasets and compared with several advanced methods, and the experimental results prove that our model has better results and more robust performance. In the future, we can consider using our model in large point cloud tasks such as semantic segmentation.

**Acknowledgements** The paper was supported by the Natural Science Foundation of Shandong Province, China (Grant No. ZR2019MF073), the Open Research Fund from Shandong Provincial Key Laboratory of Computer Network (No. SDKLCN-2018-01), the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) (Grant No. 20CX05001A),

the Major Scientific and Technological Projects of CNPC (No. ZD2019-183-008), and the Creative Research Team of Young Scholars at Universities in Shandong Province (No.2019KJN019).

## References

1. Daniel, M., Sebastian, S.: Voxnet: A 3D convolutional neural network for real-time object recognition. In: International Conference on Intelligent Robots and Systems (IROS), pp. 9–922. IEEE (2015)
2. Hang, S., Subhransu, M., Evangelos, K., Erik, L.: Multi-view convolutional neural networks for 3D shape recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 9–945 (2015)
3. Tan, Y., Jingjing, M., Junsong, Y.: Multi-view harmonized bilinear network for 3D object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 186–194 (2018)
4. Gernot, R., Ali, O., Andreas, G.: Octnet: Learning deep 3D representations at high resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 35–3577 (2017)
5. Truc, L., Ye, D.: Pointgrid: A deep network for 3D shape understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 92–9204 (2018)
6. Charles, R.Q., Hao, S., Kaichun, M., Leonidas, J.: Pointnet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6–652 (2017)
7. Charles, R.Q., Li, Y., Hao, S., Leonidas, J.: Pointnet++: deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems (NeurIPS), pp. 510–5099 (2017)
8. Yue, W., Yongbin, S., Ziwei, L., Sanjay, E.S., Michael, M.B., Justin, M.S.: Dynamic graph CNN for learning on point clouds. ACM Trans. Graph. (tog) **38**(5), 1–12, 201 (2019)
9. Yingxue, Z., Michael, R.: A graph-CNN for 3d point cloud classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6–6279. IEEE (2018)
10. Can, C., Luca Zanotti, F., Antonios, T.: Gapnet: graph attention based point neural network for exploiting local feature of point cloud. arXiv preprint arXiv: 1905.08705 (2019)
11. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7–770 (2016)
12. Martin, S., Nikos, K.: Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 37–3693 (2017)
13. Jiaxin, L., Ben, M.C., Gim Hee, L.: So-net: self-organizing network for point cloud analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 94–9397 (2018)
14. Yiru, S., Chen, F., Yaoqing, Y., Dong, T.: Mining point cloud local structures by kernel correlation and graph pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 45–4548 (2018)
15. Yizhak, B., Michael, L., Anath, F.: 3DMFV: threedimensional point cloud classification in real-time using convolutional neural networks. IEEE Robot. Autom. Lett. **3**(4), 3145–3152 (2018)
16. Yifan, X., Tianqi, F., Mingye, X., Long, Z., Yu, Q.: SpiderCNN: deep learning on point sets with parameterized convolutional filters. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–87 (2018)

17. Kuangen, Z., et al.: Linked dynamic graph CNN: Learning on point cloud via linking hierarchical features. arXiv preprint arXiv: 1904.10014 (2019)
18. Yangyan, L., Rui, B., Mingchao, S., Wei, W., Xinhuan, D., Baoquan, C.: PointCNN: convolution on x-transformed points. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 8–820 (2018)
19. Zhirong, W., Shuran, S., Aditya, K., Fisher, Y., Linguang, Z., Xiaouou, T., Jianxiong, X.: 3D shapenets: a deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 192–1912 (2015)
20. Mikaela Angelina, U., Quang-Hieu, P., Binh-Son, H., Duc Thanh, N., Sai-Kit, Y.: Revisiting point cloud classification: a new benchmark dataset and classification model on real-world data. arXiv, pp. 20–908 (2019)

# Image Detection of Peach Diseases and Pests



Qi Li<sup>1</sup>, Wenjie Sun<sup>1</sup>, Aiju Shi<sup>1</sup>, Chengmin Lei<sup>1</sup>, and Shaomin Mu<sup>1</sup>

**Abstract** Aiming at the problem of low accuracy and efficiency of artificial identification of peach diseases and pests, the RFBNet based on Kmeans++ is proposed to construct the image detector of peach diseases and pests respectively. The Kmeans++ algorithm is used to adjust the prior box size instead of manually setting the prior box size, which makes it match the size of diseases and pests better, so that small diseases and pests can get better detection results. Four kinds of disease images and five kinds of pest images were collected from peach orchard in Shandong Province to construct the data set of peach diseases and pests, and the data set was expanded by five kinds of data enhancement methods to enhance the generalization ability of the model. The experimental results show that using this algorithm to detect peach disease and pest images, the disease detection accuracy is 73.12%, and the pest detection accuracy is 94.02%, which are higher than the SSD and RFBNet.

**Keywords** Diseases detection · Pests detection · RFBNet · Kmeans++

## 1 Introduction

Peach trees are one of the three major fruit trees in China, which are planted in various regions of our country. However, they are accompanied by the occurrence of diseases and pests in every growth period of peach trees. If the disease and pest cannot be identified quickly and accurately, and then apply the appropriate medicine to the disease, the quality and quality of the peach will be seriously affected. At present, the traditional methods of diseases and pests identification mainly rely on the past experience accumulation of farmers or the guidance of relevant plant protection experts, but this method has the problems of low identification accuracy and

---

Q. Li · W. Sun · C. Lei · S. Mu

College of Information Science and Engineering, Shandong Agricultural University, Taian 271018, Shandong, China

A. Shi (✉)

College of Chemistry and Materials Science, Shandong Agricultural University, Taian 271018, Shandong, China

e-mail: [Shiaiju@sda.edu.cn](mailto:Shiaiju@sda.edu.cn)

low identification efficiency, and the limited plant protection experts cannot provide technical guidance to farmers in various regions. If the disease and pest occur without timely control, it may lead to the outbreak of diseases and pests in a large area, and then cause serious economic losses.

In recent years, with the rapid development of deep learning and artificial intelligence technology, many scholars have applied object detection technology to the detection of agricultural diseases and pests, which also provides a new method for the accurate identification of peach diseases and pests. Traditional object detection methods mainly select candidate regions by sliding window, and then extract features of candidate regions by HOG [1], LBP [2] and Haar-like [3], etc. Finally, classifier is used to classify images. However, this method has the problems of poor generalization ability of feature extraction, slow detection speed and low detection accuracy. The object detection algorithm based on deep learning has made a major breakthrough in this aspect, mainly including two-stage object detection algorithm and one-stage object detection algorithm. The two-stage object detection algorithm improves the detection accuracy but the detection speed is still slow, mainly represented by RCNN [4], Faster RCNN [5] and R-FCN [6]. The one-stage object detection algorithm completes feature extraction, image classification and bounding box regression in a network, which not only ensures the accuracy of detection, but also improves the detection speed, mainly represented by YOLO [7] and SSD [8] algorithms. Therefore, in order to obtain better detection results, this paper uses a one-stage object detection method to achieve peach tree diseases and pests detection. For this paper, the main contributions are as follows: (1) The image data set of peach tree diseases and pests was constructed. (2) Aiming at the problem that diseases and pests are difficult to detect, the RFBNet based on Kmeans++ is proposed to construct image detector of peach diseases and pests respectively, so as to realize the detection of diseases and pests with smaller volume.

## 2 Related Work

### 2.1 Agricultural Diseases Detection

In recent years, object detection based on deep learning has made some achievements in agricultural diseases recognition. Sadojevic et al. [9] modified and fine-tuned Caffenet, and retrained Softmax classifier to realize diseases detection of 13 different types of crops. Fuentes et al. [10] realized the real-time detection of 9 different types of tomato diseases and pests by using the method based on deep learning, and compared different combinations of models and feature extractors through experiments to prove the combined detection accuracy of R-FCN and ResNet-50 higher. Karthik et al. [11] used two different architectures to detect tomato diseases. The first is to use residual learning to classify tomato diseases, and the second is to add notes to the top of the residual structure. The experimental results show that the second structure has

better detection effect. Xinran et al. [12] improved Faster RCNN model to solve the problem that the disease spots on apple leaves are small and difficult to detect in complex background, using feature pyramid to increase the ability of network feature extraction, and using accurate region of interest pooling to reduce pixel loss caused by quantization operation.

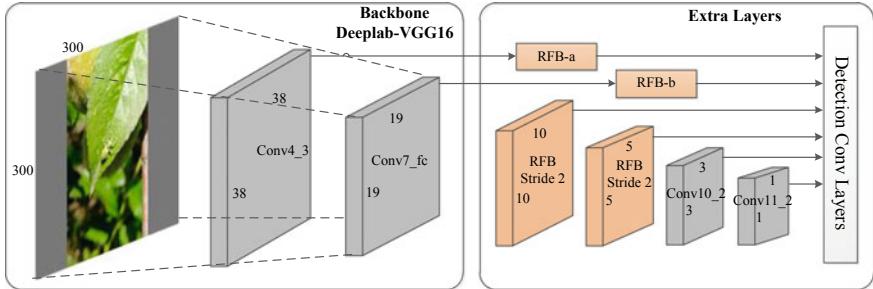
## 2.2 Agricultural Pests Detection

Ding and Taylor [13] used the convolutional neural network sliding window method to detect whether there were moths in ROI, and the experiment showed that the convolutional neural network method had a higher recall rate than LogReg algorithm. Shen et al. [14] used the detection method of Faster R-CNN to realize the detection of stored grain pests under field conditions, and the optimized network can improve the detection accuracy of small size pests. Selvaraj et al. [15] trained three different detection models with the deep learning-based migration method, and realized the detection of diseases and pests in different parts of bananas, which provided a foundation for the transplantation of the model to mobile phones. Wang et al. [16] used four object detection algorithms of Faster RCNN, SSD, YOLOV3 and Cascade RCNN to detect 24 types of pests, and the experiment showed that the number, scale and stacking adhesion of pests were the three main factors affecting the detection accuracy of pests.

## 3 Method

### 3.1 RFBNet Object Detection

RFBNet is a lightweight object detection model based on regression, which can complete image feature extraction, image classification and border regression in a network, and the detection speed is fast. Its structure is similar to the SSD model, with the VGG architecture as the backbone network, as shown in Fig. 1. The convolution layer is used to replace the original Fc6 and Fc7 layers. Four layers of network are added in the back, the first two layers are RFB modules and the last two layers are convolution layers; and RFB-a and RFB-b modules are respectively connected behind Conv4\_3 and Conv7\_fc. Inception structure and the idea of dilated convolution are used to simulate human visual perception mechanisms to achieve better detection effects. The main characteristics of RFB module are: using convolution kernels of different sizes to form a multi-branch convolution structure to simulate the size of different receptive fields; dilated convolution can obtain higher resolution image features while keeping the parameters unchanged. RFBNet takes  $300 \times 300$  images as input. In order to obtain feature information of different scales, six effective feature



**Fig. 1** RFBNet model

images are used to classify and detect objects, which improved the detection accuracy of small objects. Finally, the final detection results are obtained by non-maximum suppression method.

### 3.2 Kmeans++ Optimization Prior Box

In RFBNet [17], the location of the object is determined by learning the changes of prior box, but the number and size of prior box are manually selected, which do not match the image data set of peach diseases and pests. If the optimal size of prior box can be determined in the initial solution stage, it is easier to learn the exact location of diseases and pests in the network, and then improve the final detection accuracy. In this paper, the Kmeans++ algorithm [18] is used to adjust the size of the prior box, so as to better match the size of peach tree diseases and pests. The idea of the algorithm is to select the point with the largest distance as the initial clustering center to make the distribution of the clustering center more uniform and reduce errors. The algorithm steps are shown in Table 1.

Kmeans++ optimizes the selection method of the initial clustering center, which can obtain better clustering effect. It can be seen from Figs. 2 and 3 that with the increase of the number of cluster centers, the average IOU value increases. When the clustering center is 6, the average IOU of the disease images increases slowly. When the clustering center is 8, the average IOU of the pest images increases slowly. Therefore, the number of clustering centers of disease images is set as 6, and the number of clustering centers of pest images is set as 8. Tables 2 and 3 are the size of prior box of disease images and the size of prior box of pest images after clustering with Kmeans++, respectively.

**Table 1** Steps of Kmeans++ algorithm

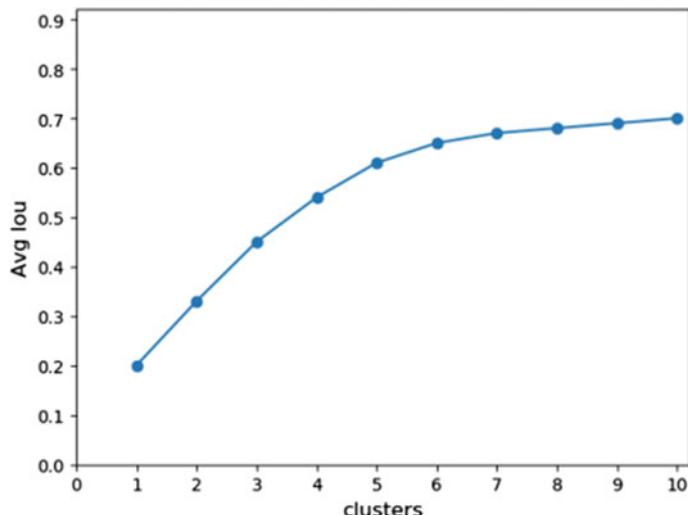
---

Algorithm: Kmeans++ clustering prior box

---

1. Randomly select one of all the target boxes of the data set as the first clustering center
  2. Calculate the distance between the input target box and the nearest clustering center  $d = 1 - \text{IOU}$ . If the distance between a certain target box and the clustering center is greater than the distance threshold, the target box will be selected as the next clustering center
  3. Repeat the above steps until  $K$  cluster centers are selected
  4. Calculate the distance between the remaining target frame and each clustering center, and which clustering center has the smallest distance will become its cluster
  5. Recalculate the cluster center of each cluster, the calculation method is  

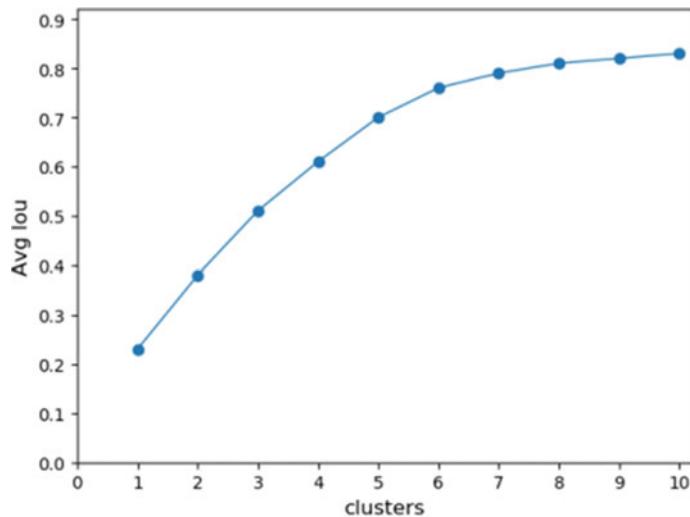
$$W'_i = \frac{1}{N_i} \sum w_i, H'_i = \frac{1}{N_i} \sum h_i$$
, where  $N_i$  is the number of target boxes in the  $i$ -th cluster
  6. Repeat steps 4 and 5, and when the variable in the cluster center is small, stop iteration to get the final cluster center
- 

**Fig. 2** Clustering center of disease images

## 4 Experiment

### 4.1 Construction of Data Set

**Data collection.** The images of peach diseases and pests were collected from peach orchards in Qingdao, Shandong Province. The images of peach tree diseases and pests with a resolution of  $2240 \times 3968$  in natural environment were obtained by random distance from different angles. A total of 2066 images of four common peach tree diseases were collected, including peach leaf curl, brown spot perforation, bacterial



**Fig. 3** Clustering center of pest images

**Table 2** Result of prior box selection of disease images

Width (pixel)	High (pixel)	Aspect ratio
35.09	26.70	1.3
18.48	11.57	1.6
119.73	157.60	0.76
66.96	76.13	0.88
11.12	6.88	1.62
5.76	3.63	1.59

**Table 3** Result of prior box selection of pest images

Width (pixel)	High (pixel)	Aspect ratio
27	22	1.23
27	47	0.57
46	60	0.77
143	165	0.87
49	40	1.22
20	33	0.61
32	33	0.97
16	19	0.84

**Table 4** Disease images of peach

Disease name	Leaf curl	Anthrax	Bacterial perforation	Brown spot perforation
Image				

**Table 5** Pest images of peach

Pest name	<i>Empoasca fabae</i>	Green weevil	<i>Monema flavesrens</i>	<i>Myzus persicae</i>
Image				
Pest name	Lacewing	<i>Halyomorpha halys</i>	Ladybug adults	Ladybug larvae
Image				

perforation and anthrax. And 2882 images of five common peach tree pests, such as *Myzus persicae*, green weevil, *Halyomorpha halys* and *Monema flavesrens*. Although lacewing and ladybug are not pests, they are natural enemies of pests and are also suitable for peach orchard pests detection. Some images of peach disease and pest collected are shown in Tables 4 and 5.

**Data Augmentation.** In order to meet the input requirements of the model, the image size is uniformly adjusted to  $300 \times 300$ . Usually, adjusting the image size will destroy its aspect ratio, which will lead to the deformation of the object in the image and affect the feature extraction. In this paper, Letterbox is used to fill the image into a square, and then scale it proportionally, so that the scaled image still keeps its original shape. Figure 4 is the original image, and Fig. 5 is the image transformed by Letterbox.

After normalizing the image size, LabelImg software was used to manually mark the images of peach diseases and pests, the position of the diseases and pests and the corresponding category name were mainly marked by drawing a box, and the marked images were stored in Pascal VOC2007 data set format. Because the collection of peach tree diseases and pests images is limited by weather and season, the collection quantity is too little, and manual labeling of data boxes will take a lot of time. Therefore, in order to enhance the generalization ability of the model, this paper uses five

**Fig. 4** Original image



**Fig. 5** Letterbox transform image



**Table 6** Image enhancement method

Original image	Cramming	Flipping	Brightness adjustment	Translation	Gaussian noise
					

**Table 7** Parameter setting

Parameter	Value
Training set:test set	9:1
Initial learning rate	0.0001
Number of iterations	100
Batch_size	32
Optimizer	Adam

kinds of data augmentation: “cramming”, translation, brightness adjustment, flipping and adding Gaussian noise. After the data augmentation operation, the number of disease images reached 12,452, and the number of pest images reached 15,367. Taking disease images as an example, the image enhancement methods adopted in this paper are shown in Table 6.

## 4.2 Experimental Environment and Parameter Setting

The experimental environment in this article uses Intel(R) Core(TM) i5-6600U CPU @3.30 GHz processor, memory is 16 GB, graphics card model is NVIDIA GeForce GTX 1080Ti, operating system is Windows 10, and Python3.6 is used as the programming language. The deep learning framework is Keras 2.1.5.

The parameter setting of the model is shown in Table 7.

## 4.3 Evaluation Index

In this paper, the average precision (AP) and mean average precision (mAP) of peach diseases and pests are used as evaluation indexes. The calculation method is shown in formula (2) and (3), and the mAP is finally determined by precision, as shown in formula (1).

**Table 8** Image detection results of peach diseases under different models

Model	Leaf curl	Anthrax	Bacterial perforation	Brown spot perforation	mAP(%)
	AP (%)				
SSD	72.46	70.47	65.45	68.52	69.25
RFBNet	74.23	72.41	67.88	70.82	71.31
Improved RFBNet	<b>76.58</b>	<b>74.32</b>	<b>69.23</b>	<b>72.34</b>	<b>73.12</b>

$$precision = \frac{TP}{TP + FP} \quad (1)$$

$$AP = \frac{\sum_c precision}{C} \quad (2)$$

$$mAP = \frac{\sum_{n=1}^N AP(n)}{N} \quad (3)$$

#### 4.4 Analysis of Experimental Results

In order to verify the effectiveness of the algorithm in this paper, it is compared with the SSD and RFBNet algorithms. The peach disease images detection results are shown in Table 8, and the peach pest images detection results are shown in Table 9.

It can be seen from Table 8 that the mAP of peach diseases detected by SSD is 69.25%, and that detected by RFBNet is 71.31%, after optimizing the prior box in the network by Kmeans++ algorithm, its size is more in line with the size of peach diseases, which is 3.87% and 1.81% higher than the detection accuracy of the first two models respectively, and the detection accuracy of each kind of diseases is improved.

It can be seen from Table 9, the mAP of SSD for the detection of peach pests is 87.46%, the mAP of RFBNET for the detection of peach pests is 92.63%, and the mAP of RFBNet after adjusting the prior box size is 94.02%, which is 6.56% and 1.39% higher than the previous two models respectively. In addition, the detection accuracy of each kind of pests is improved, and the AP value of *Halyomorpha halys* reached 99.51%, indicating that better detection effect could be achieved after the optimization of the prior box.

**Table 9** Image detection results of peach pests under different models

Model	<i>Empoasca fabae</i>	Green weevil	<i>Monema flavaescens</i>	<i>Myzus persicae</i>	Lacewing	<i>Halyomorpha halys</i>	Ladybug adults	Ladybug larvae	mAP (%)
	AP (%)								
SSD	81.75	90.82	92.55	67.30	88.36	93.13	92.21	93.26	87.46
RFBNet	85.46	94.14	97.83	79.16	91.25	98.74	96.95	97.54	92.63
Improved RFBNet	<b>87.69</b>	<b>96.31</b>	<b>99.07</b>	<b>80.41</b>	<b>92.83</b>	<b>99.51</b>	<b>99.16</b>	<b>97.15</b>	<b>94.02</b>

## 5 Conclusion

Aiming at the detection problem of peach diseases and pests, this paper proposes to optimize the prior box in RFBNet model by using Kmeans++ algorithm, so as to obtain better detection results. The dilated convolution in RFBNet can guarantee the parameters without changing the resolution of the image, and realize the feature fusion of different convolution outputs in the RFB module, which has a better detection effect on small targets. In order to set the size of the prior box in accordance with peach tree diseases and pests, Kmeans++ algorithm is used to adjust the size of the prior box. Finally, the proposed algorithm is compared with SSD and RFBNet on peach tree diseases and pests images data set. The results show that the proposed algorithm has higher detection accuracy. In the next step, we will continue to expand the data set of peach tree diseases and pests to realize the detection of more kinds of peach tree diseases and pests. The structure design of the model will be improved, and a special model for image detection of peach tree diseases and pests will be studied and designed.

**Acknowledgements** This work was partially supported by First Class Discipline Funding of Shandong Agricultural University.

## References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893. IEEE, Los Alamitos, CA (2005)
2. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of 12th International Conference on Pattern Recognition, pp. 582–585. IEEE, Los Alamitos, CA (1994)
3. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proceedings of International Conference on Image Processing, pp. I-I. IEEE, Piscataway, NJ (2002)
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. IEEE, Los Alamitos, CA (2014)
5. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2015)
6. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. *Adv. Neural. Inf. Process. Syst.* **29**, 379–387 (2016)
7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. IEEE, Los Alamitos, CA (2016)
8. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham, Switzerland (2016)
9. Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., Stefanovic, D.: Deep neural networks based recognition of plant diseases by leaf image classification. *Comput. Intell. Neurosci.* **2016**, 1–11 (2016)

10. Fuentes, A., Yoon, S., Kim, S.C., Park, D.S.: A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* **17**(9), 2022 (2017)
11. Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., Menaka, R.: Attention embedded residual CNN for disease detection in tomato leaves. *Appl. Soft Comput.* **86**, 105933 (2020)
12. Xinran, L., Shuqin, L., Bin, L.: Detection of apple leaf diseases based on improved Faster R\_CNN. *Comput. Eng.*, 1–7 (2020)
13. Ding, W., Taylor, G.: Automatic moth detection from trap images for pest management. *Comput. Electron. Agric.* **123**, 17–28 (2016)
14. Shen, Y., Zhou, H., Li, J., Jian, F., Jayas, D.S.: Detection of stored-grain insects using deep learning. *Comput. Electron. Agric.* **145**, 319–325 (2018)
15. Selvaraj, M.G., Vergara, A., Ruiz, H., Safari, N., Blomme, G.: AI-powered banana diseases and pest detection. *Plant Methods* **15**(1), 92 (2019)
16. Wang, Q.J., Zhang, S.Y., Dong, S.F., Zhang, G.C., Yang, J., Li, R., Wang, H.Q.: Pest24: a large-scale very small object data set of agricultural pests for multi-target detection. *Comput. Electron. Agric.* **175**, 105585 (2020)
17. Liu, S., Huang, D.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 385–400. Springer, Cham (2018)
18. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp.1027–1035. SIAM, New Orleans (2007)

# Automatic Classification of Tongue Shape Based on Improved Analytic Hierarchy Process



Shanshan Gao , Liqian Zhang , Menghang Li , and Wenhan Dou

**Abstract** The shape of the tongue is one of the important characteristics of the tongue, and it is related to the pathological degree of viscera and organs. In TCM (traditional Chinese medicine) tongue diagnosis, the shape of the tongue is subjectively judged by the TCM physician's observation of the tongue body and comparison with the normal tongue shape. However, it is difficult to establish a standard for the normal tongue shape. Moreover, the skewed tongue body will bring a certain error to the classification of the tongue shape. Therefore, this paper proposes a tongue correction method, and the classification of tongue shape is realized on the basis of the proposed method. First, based on the symmetry characteristics of the tongue, this paper uses the Harris corner detection method to extract the tongue's central axis to correct the skewed tongue. Second, seven common tongue shapes were classified by Analytic Hierarchy Process (AHP) based on the tongue length and area related features. The tongue correction experimental results prove the effectiveness of the correction method proposed in this paper, and the tongue classification experimental results prove that classification effect of the tongue shape is more accurate.

**Keywords** Tongue correction · Corner detection · Tongue classification · Analytic hierarchy process

---

S. Gao · L. Zhang · M. Li · W. Dou  
Shandong University of Finance and Economics, Jinan 250014, China

S. Gao  
Shandong Provincial Key Laboratory of Digital Media Technology, Jinan 250014, China  
Shandong China-U.S. Digital Media International Cooperation Research Center, Jinan 250014, China

## 1 Introduction

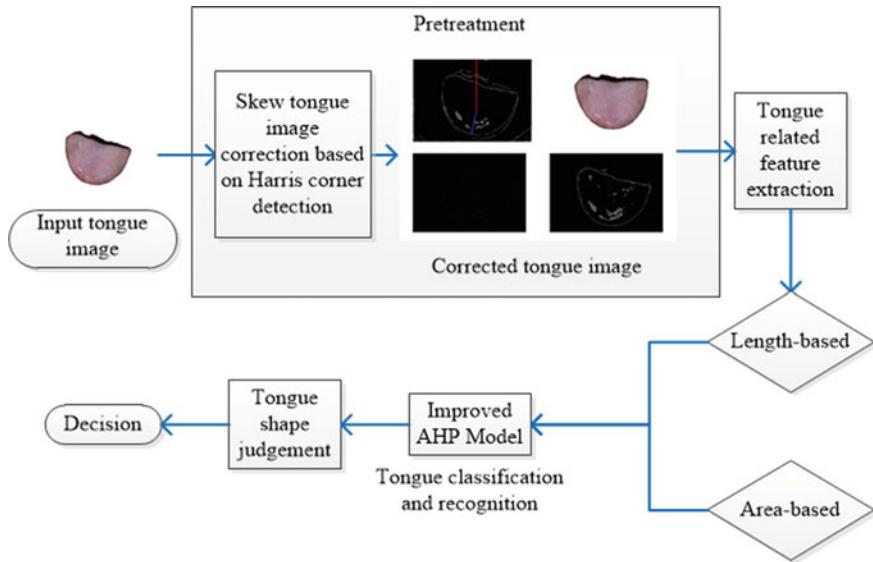
Tongue diagnosis is a very characteristic diagnosis and treatment technique of traditional Chinese medicine. In recent years, researchers have tried to use computer technology to identify and diagnose tongue images. Many computerized tongue diagnostic instruments have been developed successively and good research results have been obtained. However, the research of computerized tongue image diagnosis mainly focuses on the color or texture characteristics of the tongue image, and there are relatively few studies on the shape of the tongue.

The shape of the tongue is also one of the important characteristics of the tongue picture. Some clinical reports [1] point out that there is an inseparable connection between specific tongue shapes and clinical medical diseases. For example, a patient with a round tongue may have gastritis, and a patient with a rectangular tongue may have coronary heart disease or portal hypertension [2]. Therefore, the shape of the tongue has important clinical value in the diagnosis of diseases and syndromes, and the research on the classification algorithm of the tongue shape is necessary for the informationization of the tongue diagnosis of Chinese medicine.

At present, there are few studies on the classification and quantification of TCM syndromes, for the following reasons. One is that when the tongue image is collected, the patient's strength is inconsistent during the tongue extension process, which is prone to the phenomenon of tongue shaking, which leads to the acquisition of a skewed tongue. This is not conducive to the precise positioning of the local characteristics of the tongue shape such as tongue length and tongue width, which affects the accuracy of the computer's recognition of the tongue shape. The second is that the subjective analysis of tongue shape using computer should be defined as a measurable machine representation. Third, the lack of samples with different tongue shapes greatly limits the research based on methods such as deep learning.

In this paper, a new method for the correction and classification of the skew tongue is proposed. In our method, the center axis of the tongue was determined by Harris corner detection method to correct the skewed tongue. Based on the precise contour curve of the tongue edge, the hierarchical structure model was constructed through decision support tools using five tongue-related features based on length and area, and the classification of seven common tongue shapes such as square tongue and rectangular tongue was realized. The procedure of the tongue classification method in this paper is shown in Fig. 1.

The rest of this article is organized as follows. In the second section, the related work of this paper is described. In the third section, the tongue body alignment algorithm based on Harris corner detection and the tongue shape classification algorithm based on AHP are introduced respectively. The fourth part shows the results of tongue correction and tongue shape classification experiments. Finally, we list the conclusions of the whole work.



**Fig. 1** Tongue classification method process

## 2 Related Work

### 2.1 Tongue Correction

As described in the first section, before the classification of tongue shape, the deflected tongue image needs to be corrected. In 2003, Wei [3] combined the positioning of the corners of the mouth and the comparison of radius-angle diagrams to determine the central axis of the tongue. The tongue skew index is calculated based on the ratio of the area difference on both sides of the central axis to the tongue image area, the position of the connecting segment of the corner of the mouth and the center point. Subsequently, Zhu [4] obtained the axis of symmetry by calculating the distance between the tongue weight center and the symmetry point, extracted the position of the corners of the mouth with the histogram of the variation tone component, and calculated the skew coefficient by using the parameters of the line between the central axis and the corners of the mouth. However, the above two methods only reflect the degree of tongue skew, and do not perform tongue correction. In 2006, Yang [5] used spline interpolation to perform least squares polynomial curve fitting on the sampled edge points after extracting the tongue boundary line. After obtaining the tongue corner points and rectangular coordinates through the curvature of the curve, combined with the tongue width Obtain the central axis symmetry line of the tongue profile. Huang [6] used three geometric criteria based on the length, area and angle of the tongue to correct the skewed tongue shape, Tayo [7] took advantage of mirror

symmetry and used the method of axis of symmetry to correct the deflected tongue. The above methods all better reduce the skew degree of the tongue shape.

## 2.2 Tongue Classification

Some researchers have conducted research on the classification of tongue shapes after correcting the oblique tongue. Wu [8] and others first introduced the analytic hierarchy process to the field of pattern recognition and applied it to tongue classification. Xu [9] used morphological image processing methods to study and judge fat tongue and thin tongue. The paper also verified and analyzed the shape of the tongue corresponded to diseases such as diabetes, hypertension, and chronic gastritis. Huang [6] determined the classification of tongue shape based on the geometric features of tongue length, area and Angle, combined with the fuzzy fusion framework and seven AHP modules. Zhu [10] improved the defect of central axis extracted by area symmetry in literature [3]. Based on the shape symmetry of the tongue body, the Angle analysis method and central axis extraction method were combined to analyze the degree and direction of tongue skew, which improved the analysis accuracy. Tayo [7] used geometric features such as tongue area and bisector, algebraic features such as contour curves and curve error of tongue, and used support vector machines and multilayer perceptual neural networks to classify five tongue shapes. Literature [11] defines three different tongue shapes based on geometrical features: circle, square and triangle, and then uses different classifiers to automatically recognize and classify different tongue shapes. This method has the highest classification accuracy for round tongues, reaching 88.65%.

## 2.3 Analysis of Tongue Pathology

Some scholars have also analyzed and studied the health and pathology of tongue shape. Literature [12] proposes a feature analysis model that can learn advanced features from deep tongue images. The model learns useful features from clinical data in an unsupervised manner, and uses the acquired features to classify patients' physical conditions as healthy or abnormal using supervised machine learning techniques. In paper [13], HOG feature of tongue image was extracted and used as input of combinational classifier KNN and SVM, and distance measurement learning method was used to predict patients' pathological characteristics. Literature [14] first segmented the tongue image, and divided the segmented image into 6 partitions. In each section, the color features and number of key points of the tongue body were extracted, and the self-organizing feature mapping algorithm was used to classify the tongue image to determine whether the tongue shape belonged to diabetic tongue shape.

### 3 Method

#### 3.1 Correction of Skewed Tongue Based on Harris Corner Detection

In general, the central axis of the tongue extends along the direction of the tongue, and the shape of the tongue is distributed approximately symmetrically on both sides of the central axis. Based on this idea, this paper proposes a method of using symmetric central axis to correct the deflected tongue.

By calculating the tongue corner points and the center of gravity point, the tongue symmetry axis is obtained, and then the tongue body could be corrected according to the angle between the tongue symmetry axis and the mid-perpendicular line to assist the subsequent tongue image classification.

This paper uses Harris corner detection algorithm [15] to determine corners, searches for n pixels on the left and right sides of the intermediate data in the tongue image, and performs y-axis projection processing on the pixels. Set the threshold and calculate the mean coordinate of the pixel point, and calculate the nearest corner point by the distance function as the tongue corner points.

The center of gravity of the tongue image can be expressed as:

$$x_1 = \frac{\sum p_i x_i}{\sum p_i}, \quad y_1 = \frac{\sum p_i y_i}{\sum p_i} \quad (1)$$

In the formula,  $(x_i, y_i)$  is the coordinate value of the pixel, and  $p_i$  is the corresponding pixel value.

The process of the algorithm is shown in Algorithm 1:

*Algorithm 1*

*Input: Tongue image I*

*Output: Tongue image S*

*Step1: Obtain the corner points in I.*

*Step2: Determine the location of the tip of the tongue  $m_1(x_1, y_1)$  in I.*

*Step3: Determine the center of gravity of the tongue image pixel  $n_2(x_2, y_2)$  according to (1).*

*Step4: Determine the tongue symmetry axis of I.*

*Step5: Calculate the angle between the symmetrical axis of the tongue and the vertical line.*

*Step6: Rotate the tongue image I and save as S.*

### 3.2 Classification of Tongue Shape Based on Analytic Hierarchy Process

Figure 2 shows seven common tongue shapes. We describe these common tongue features from the perspectives of the length and width of the tongue, the tip of the tongue, the distance between the tongue base and the center of gravity, etc. The corresponding characteristics of the different tongue shapes are shown in Table 1.

**Extraction of Tongue Feature Value.** Assume that the area of the tongue body is  $Tongue\_ratio$ , the maximum and minimum values of the tongue contour points along the X-axis are  $x_{\max}$  and  $x_{\min}$  respectively, and the maximum and minimum values along the Y-axis are  $y_{\max}$  and  $y_{\min}$  respectively, then the length, width and radius of the tongue are  $Length = y_{\max} - y_{\min}$ ,  $Width = x_{\max} - x_{\min}$  and  $r = \frac{Width}{2}$  respectively.

#### Area correlation characteristics

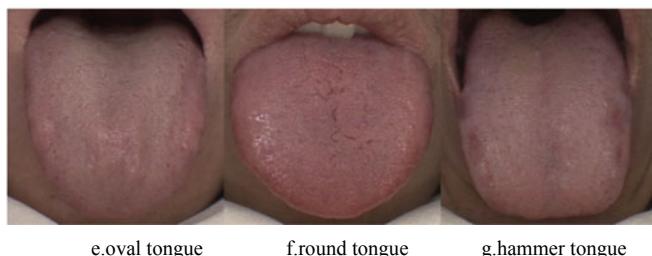
- (1) Round area ratio  $Round\_ratio$ : Describe the degree to which the tongue shape resembles a circular tongue. Set the ratio of circular area as, then

$$Round\_ratio = \frac{Tongue\_area}{r^2} \quad (2)$$

- (2) Square area ratio: Describe how similar the tongue shape is to the square tongue shape. Suppose the ratio of square area is  $Square\_ratio$ , then



a. square tongue      b. rectangular tongue      c. sharp triangular tongue      d. blunt triangular tongue



e. oval tongue      f. round tongue      g. hammer tongue

**Fig. 2** 7 common tongue shapes

$$\text{Square\_ratio} = \frac{\text{Tongue\_area}}{r^2} \quad (3)$$

- (3) Triangle area ratio: Describe the degree of similarity between the tongue shape and the triangular tongue. The areas of the sharp and blunt triangular tongues are close to that of a regular triangle. Let the base width of the triangle be  $(x_{\max} - x_{\min})$ , the base length of the triangle be  $(y_{\max} - y_{\min})$ , and the triangle area ratio be *Triangle\_ratio*, then

$$\text{Triangle\_ratio} = \frac{1}{2} \cdot \frac{\text{Tongue\_area}}{(x_{\max} - x_{\min}) \cdot (y_{\max} - y_{\min})} \quad (4)$$

#### *Length related features*

- (1) Length to width ratio: Describe the ratio of the overall length of the tongue. Let the length-width ratio be *Lw\_ratio*, then

$$\text{Lw\_ratio} = \frac{\text{Length}}{\text{Width}} \quad (5)$$

- (2) Center distance ratio: Describe the position relation between these two points in the tongue. Let the center point of the tongue body be  $p_i(x_i, y_i)$ , its coordinate be  $(\frac{x_{\max} + x_{\min}}{2}, \frac{y_{\max} + y_{\min}}{2})$ , the centrifugal point of the tongue body be  $p_a = (x_a, y_a)$ , and  $x_a$  be the midpoint between  $x_{\max}$  and  $x_{\min}$  in the tongue. Set the center distance ratio as *Cp\_ratio*, then.

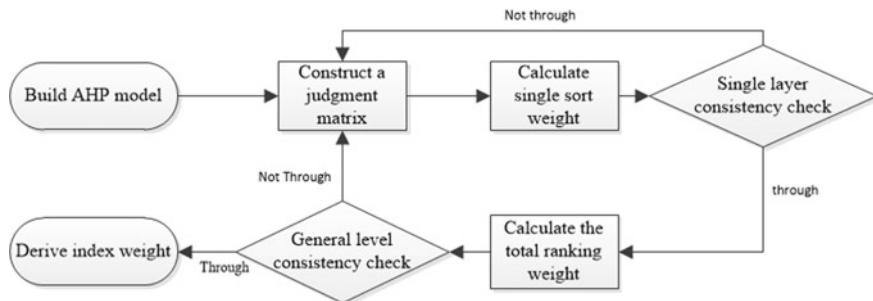
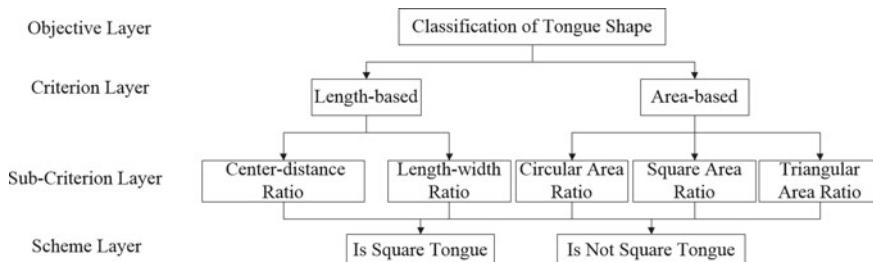
$$\text{Cp\_ratio} = \frac{y_a - y_i}{\text{Length}} \quad (6)$$

The different tongue shapes corresponding to different characteristic parameters are shown in Table 1.

**Tongue Classification Based on AHP Algorithm.** In this section, we will use five tongue-shaped related feature values based on the length and area of the tongue for the analysis of tongue-shaped classification. Since the calculated tongue-shaped related features are associated with multiple tongue-shaped categories, but different tongue shapes have different degrees of association with the same feature, a decision-making method is needed to comprehensively analyze different types of tongue-shaped features. Determine the most influential tongue-shaped related features corresponding to different tongue-shaped categories. This paper proposes the use of analytic hierarchy process to classify 7 common tongue shapes such as square, rectangle, and circle. The overall steps of the analytic hierarchy process are shown in Fig. 3. Figure 4 shows the establishment of a hierarchical structure model with a square tongue as an example.

**Table 1** Characteristics parameters of different tongue shape

	Parameter range	Tongue shape
Round area ratio	$Round\_ratio = \pi$	Round
Square area ratio	$Square\_ratio = 4$	Square
Triangle area ratio	$Triangle\_ratio \approx 1$	Sharp triangle; blunt triangle
	$Triangle\_ratio \approx 2$	Rectangle
	$1 < Triangle\_ratio < 2$	Oval; round; hammer
Aspect ratio	$Lw\_ratio > 1$	Rectangle; oval; hammer
	$Lw\_ratio \approx 1$	Round; square
	$Lw\_ratio > 1$	Sharp triangle
	$Lw\_ratio < 1$	Blunt triangle
Center distance ratio	$Cp\_ratio > 0$	Rectangle; Sharp triangle; Blunt triangle
	$Cp\_ratio < 0$	Hammer
	$Cp\_ratio = 0$	Round; Square; Oval

**Fig. 3** Hierarchical analysis process**Fig. 4** Analytic hierarchy structure diagram

**Improvement of Tongue Classification Method.** Applying the AHP algorithm to the problem of tongue shape classification, the length-related features and area-related features have equal importance measures for the seven tongue shapes mentioned in this paper, and the elements in the sub-criteria layer have different importance for different tongue shapes. Therefore, based on this characteristic, we adjust the classification of tongue shape as follows to reduce the cumbersomeness of the AHP method and improve the accuracy of tongue shape classification.

*Adjust the tongue shape judgment sequence in the improved AHP method.* Generally speaking, in the AHP analysis model we construct, the tongue shape is judged in the order of rectangle, square, blunt triangle, sharp triangle, circle, ellipse, and hammer. Through the characteristics of tongue-shaped related characteristic parameters, we make several adjustments:

- (1) Since the rectangular tongue, the oval tongue and the hammer-shaped tongue have the same aspect ratio element characteristics, when the tongue is judged as a non-rectangular tongue, the judgment of the oval tongue and the hammer-shaped tongue shall be carried out in order;
- (2) Since the two elements based on the length-related characteristics have the same parameter characteristics for the square tongue and the round tongue, when it is judged as a non-square tongue, the round tongue is judged first;
- (3) Since the blunt triangular tongue and the sharp triangular tongue have the same characteristics based on the area-related characteristics and the center distance ratio element, but the aspect ratio element has the opposite characteristics, when it is judged as a non-sharp triangular tongue, the judgment of the blunt triangular tongue is first performed;
- (4) Since the center distance ratio of the elliptical tongue, the hammer-shaped tongue and the sharp triangle tongue is different from the characteristics of the elements, when the tongue is judged as a non-elliptical tongue, the hammer-shaped tongue and the sharp triangular tongue are judged in order.

*Adjust the conditions of tongue shape determination in the improved AHP method.* We found that when the characteristic parameters of the tongue meet certain conditions, the shape of the tongue can be directly determined, without using other characteristic elements for the determination of the AHP method. The tongue shape and corresponding condition characteristics that can be directly determined are shown in Table 2.

**Table 2** Characteristic conditions and tongue shape

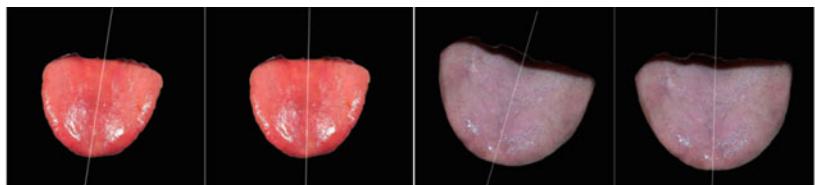
Characteristic conditions	Determinable tongue shape
$Round\_ratio > 4$ and $k_1 < Cp\_ratio < k_2$	Square
$Cp\_ratio < 0$ and $Lw\_ratio > 1$	Hammer
$Lw\_ratio < 1$ and $Cp\_ratio > k_3$	Blunt triangle

## 4 Experiments and Results

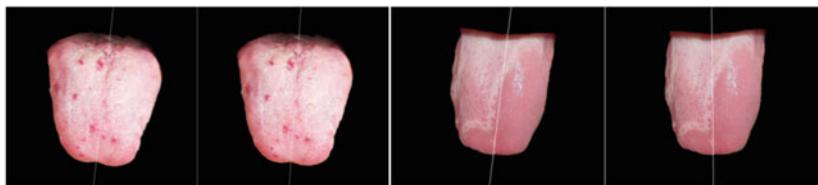
### 4.1 Skewed Tongue Image Correction

We use the Harris corner detection algorithm described in this article to obtain the central axis of the tongue and realize the correction of the skewed tongue. The algorithm experiment is implemented by MatLab programming, running on Intel(R) Core(TM) i7-8750H processor and 16 GB of memory. The results of partial skew tongue correction are shown in Fig. 5. Figures (a)–(f) respectively show images 1–6 and their corrected images. The dotted line in the figure is the central axis of tongue symmetry. It can be seen from the image that the correction results are more effective for tongues with small or large skew.

For the tongue correction method proposed in this paper, we use the mutual information in image registration to evaluate its accuracy. In image registration,



(a) Tongue image 1 and corrected tongue image (b) Tongue image 2 and corrected tongue image



(c) Tongue image 3 and corrected tongue image (d) Tongue image 4 and corrected tongue image



(e) Tongue image 5 and corrected tongue image (f) Tongue image 6 and corrected tongue image

**Fig. 5** Partial tongue body alignment results

**Table 3** Tongue's MI values

	MI <sub>1</sub>	MI <sub>2</sub>	MI <sub>1</sub> – MI <sub>2</sub>
Image 1	1.2558	1.2327	0.0261
Image 2	0.5346	0.5056	0.029
Image 3	1.3497	1.2472	0.1025
Image 4	0.9831	1.219	-0.2359
Image 5	1.1135	1.1032	0.0103
Image 6	1.2906	1.3596	0.069

the mutual information of two images reflects the degree of mutual inclusion of information between them through their entropy and joint entropy. For images R and F, the mutual information is expressed as

$$MI(R, F) = H(R) + H(F) - H(R, F) \quad (7)$$

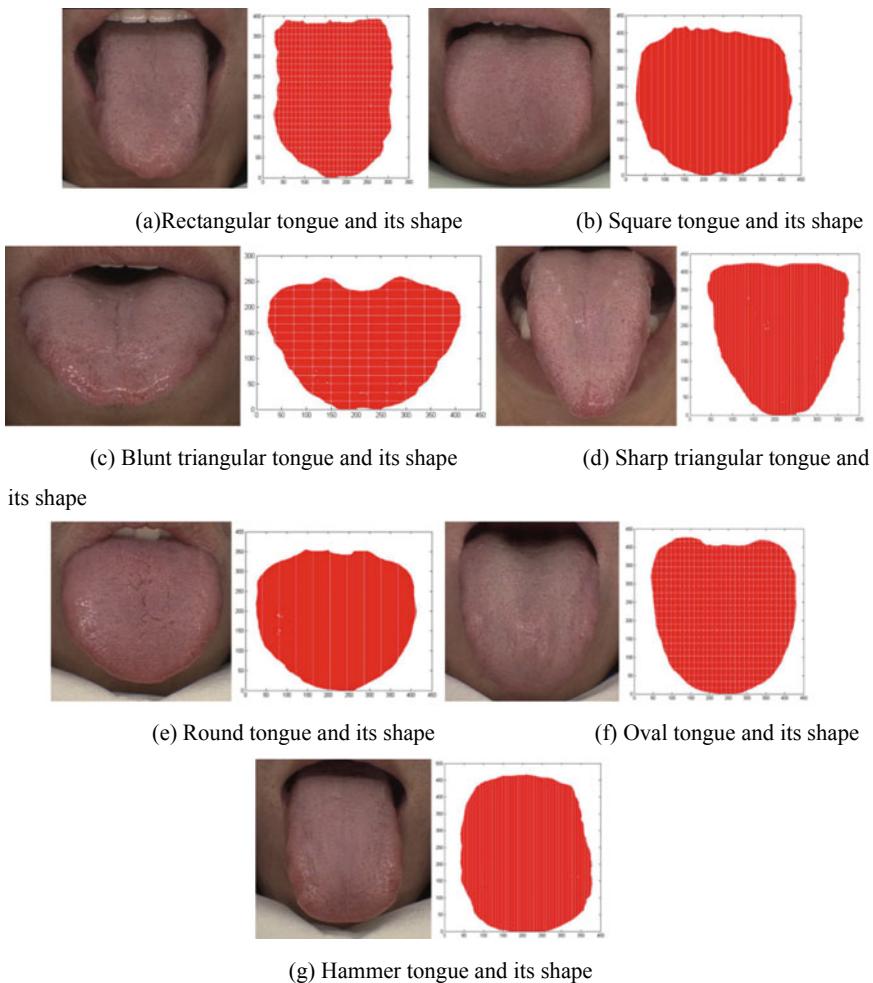
We obtain MI<sub>1</sub> and MI<sub>2</sub> by calculating the mutual information between the original tongue image and the tongue image corrected by the method proposed in this paper and the original tongue image and the artificially corrected tongue image separately, and then the error between MI<sub>1</sub> and MI<sub>2</sub> is calculated. The results are shown in Table 3. According to the data in the table, the mutual information MI value between the artificially corrected tongue image and the tongue image corrected by this method is approximately the same, indicating that the artificially corrected tongue image can basically be consistent with the tongue image corrected by this method.

## 4.2 Tongue Classification

Under the guidance of Chinese medicine experts, we selected 355 tongue image samples in the tongue image database. Since the ellipse is a normal tongue shape, its proportion in the tongue image sample is higher than that of other tongue shapes. In order to better obtain the position of the key feature points of the tongue shape, we transform the original tongue image into the tongue shape in the coordinate system, as shown in Fig. 6. We compared the results of tongue classification with the assessment of TCM experts. The classification of tongue images is shown in Table 4.

## 5 Conclusion

In this paper, based on five geometric features of tongue body, such as area and length, the analytic hierarchy model (AHP) with relatively simple structure is applied to classify tongue shape. Before extracting the tongue features, we also proposed a



**Fig. 6** Tongue shape and corresponding renderings

**Table 4** Classification of tongue images

	Number	Correctly classified samples	Accuracy (%)
Rectangular	27	23	85.20
Square	37	32	86.50
Blunt triangular	68	63	92.60
Sharp triangular	70	64	91.40
Round	55	51	92.70
Oval	74	68	91.90
Hammer	24	20	83.30

correction method of the skew tongue based on Harris corner detection, which used the symmetric central axis of the tongue body to correct the deflection image. We use the mutual information value of the image to prove the effectiveness of the tongue correction method in this paper, and compare the tongue shape classification result with the assessment of the TCM experts, it has a good accuracy rate. In future research, it is necessary to further expand the tongue image data set, increase the number of typical tongue samples, consider using deep learning methods to study the classification of tongue shapes, and improve the speed and accuracy of tongue shape classification.

## References

1. LiYou, S., et al.: Discussion on the objective research of tongue diagnosis using computer image recognition technology. *J. Anhui Coll. Traditional Chin. Med.* **5**(4), 5–7 (1986)
2. Bo, H., et al.: Tongue shape classification by geometric features. *Inf. Sci.* **180**(2), 312–324 (2010)
3. Baoguo, W., et al.: Automatic analysis of tongue deviation. *Comput. Eng. Appl.* **25**, 22–25 (2003)
4. Mingfeng, Z., et al.: A new method of automatic analysis of tongue deviation using self-correction. *J. Biomed. Eng.* **29**(1), 152–156 (2012)
5. Zhaohui, Y., et al.: A new tongue skew correction algorithm. In: 1st National Conference on Diagnosis of Integrated Traditional Chinese and Western Medicine, pp. 37–44 (2006)
6. Bo, H., et al.: Tongue Feature Analysis and Symptom Diagnosis Classification. Harbin Institute of Technology (2009)
7. Tayo, O., et al.: Features for automated tongue image shape classification. In: IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp. 273–279 (2012)
8. Jinsong, W., et al.: Research on tongue shape judgment based on AHP. In: 1st National Conference on Diagnosis of Integrated Traditional Chinese and Western Medicine, pp. 43–50 (2006)
9. Jiatuo, X., et al.: A diagnostic method based on tongue imaging morphology. In: 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 2613–2616 (2008)
10. Mingfeng, Z., et al.: A novel approach for automatic tongue deviation analysis with auto-correction. In: International Conference on Advanced Materials and Computer Science, Key Engineering Materials, vol. 474–476, pp. 69–74. Trans Tech, Switzerland (2011)
11. Devi, G.U., et al.: An analysis of tongue shape to identify diseases by using supervised learning techniques. In: International Conference on Information Communication and Embedded Systems (2017)
12. Dan, M., et al.: A deep tongue image features analysis model for medical application. In: IEEE International Conference on Bioinformatics and Biomedicine, pp. 1918–1922 (2016)
13. Jie, D., et al.: Classification of tongue images based on doublet SVM. In: 1st International Symposium on System and Software Reliability, pp. 77–81 (2016)
14. Srividhya, E., et al.: Diagnosis of diabetes by tongue analysis. In: International Conference on Computational Intelligence and Knowledge Economy, pp. 217–222 (2019)
15. Harris, C., et al.: A combined corner and edge detector. In: 4th Alvey Vision Conference, Manchester, pp. 147–151 (1988)

# Design of Visualization System for Digitalization of the Discrete Manufacturing Industry



Jianguo Yan<sup>ID</sup>, Yu Zhao<sup>ID</sup>, Wei Wang<sup>ID</sup>, Fangfu Xu<sup>ID</sup>, Wei Zhu<sup>ID</sup>, and Lili Jin<sup>ID</sup>

**Abstract** This paper is mainly based on the author's several years of experience in enterprise digitalization projects, and designs a set of data visualization systems for the production digital scene of discrete manufacturing, aiming to help discrete manufacturing practitioners quickly realize a data visualization system that meets management requirements. In addition to the characteristics of general data visualization system, this system also makes targeted design and optimization for the characteristics of production planning of the discrete manufacturing industry, such as multiple varieties and small batches, relatively frequent changes of production plan, and multiple types of work involved, especially for real-time data and dynamic monitoring, some optimization algorithms have been made. The design plan includes, but not limited to the overall system architecture, functional module division, innovative functional design, partial implementation contents and several sample effects. And based on this design principle, the author has successfully implemented a visualization system for many local companies and received good feedback.

**Keywords** Data visualization system · Digitization of production · Software system · System design

## 1 The Discrete Manufacturing

The manufacturing industry can be divided into process manufacturing and discrete manufacturing according to the organizational characteristics of the production process. In process manufacturing, the materials (raw materials) continue to go through the processing equipment, resulting in changes in the morphological and chemical properties of the materials, and get the finished products ultimately. For example, pharmaceuticals industry and chemicals industry are typical process manufacturing industries. Comparably, the process of discrete manufacturing is more complex. Generally, discretely manufacturing products are composed of multiple

---

J. Yan (✉) · Y. Zhao · W. Wang · F. Xu · W. Zhu · L. Jin

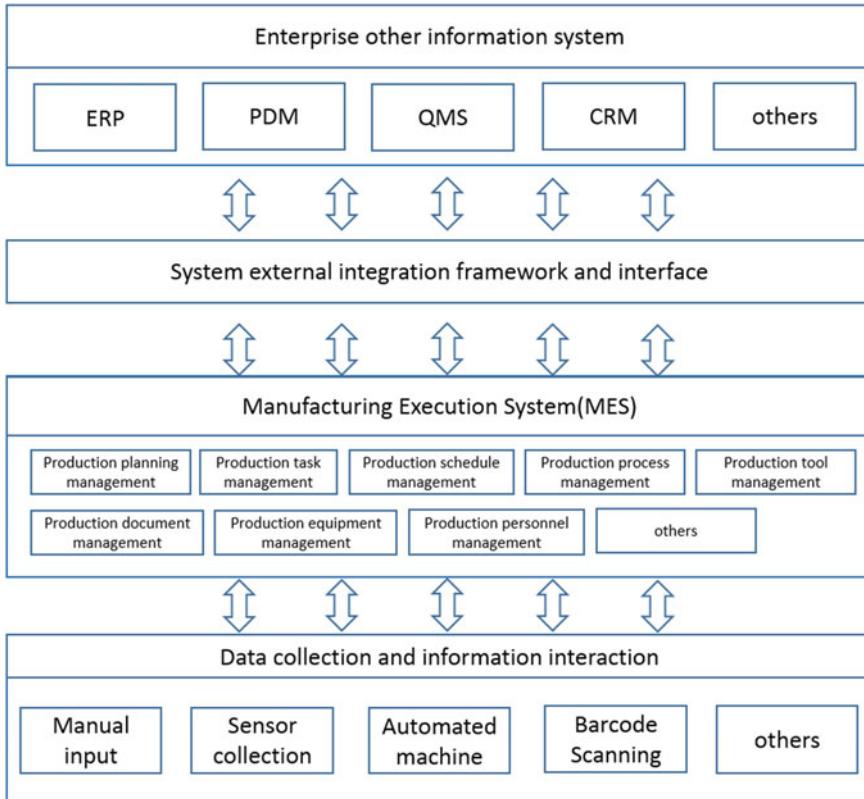
Research Institute of Mechanical & Electrical Taizhou Research Institute of Zhejiang University, Taizhou, Zhejiang, China

parts, each component may have independent processing and assembly processes, and there could exist sequential relationships between these processes. Furthermore, these parts need to be assembled after processing process has been completed to get the finished products. For instance, auto parts, electronics manufacturing, plastic production are discrete manufacturing. These characteristics often lead to more complex production management in discrete manufacturing. The production of discrete manufacturing has some characteristics as well as problems: (1) The degree of dependence on production (semi-)automation is relatively high. As China's labor costs keep rising and automation technology continues to develop, the concept of replacing labor with machines in discrete manufacturing has now become a general consensus. (2) Product bill of material (BOM) is critical. Since the finished product of discrete manufacturing is assembled from various parts, and the BOM is fixed for a specific finished product, which is also one of the most critical basic data for production management. (3) The process flow of discrete manufacturing is more complex and diverse. The production process of discrete manufacturing has the characteristics of multitasking, and also multiple devices are producing simultaneously. There are even scenarios where one device produces different processes, and the same process is produced on different devices. (4) Production plan changes frequently. In fact, discrete manufacturing will start only when there comes an order, multi-variety production, small batch production and single-piece production are tricky issues as well. Thus the production plan is comparably difficult to develop and execute. The traditional manufacturing model cannot meet the changes in new markets and new environments. It is necessary to realize the transformation from mass manufacturing to mass customization [1]. At present, the industry is able to solve these problems in stages through digital transformation of discrete manufacturing workshops, partially.

## 2 Production Digitalization

Production digitalization mainly refers to the use of various advanced technologies such as automation, computers, big data, and the Internet of Things, to do information collection, processing, reorganization, optimization of workshop production resources (devices, production materials, personnel, production processes, etc.), to provide management-related application functions of production site. Besides, it would integrate the existing information systems and display all kinds of production-related data, and results in the form of reports, charts, data kanbans, etc., thereby to enhance the transparency of workshop production process, realizing paperless office, achieving agile production information transmission and accurate production decision making as well. Use big data technology to analyze and optimize data, which will eventually realize automated, intelligent, and customized intelligent manufacturing [2]. The overall framework for the digitizing production in discrete manufacturing is shown in Fig. 1.

In the digital transformation and upgrading of the discrete manufacturing industry, the system will collect massive amounts of data. How to correctly and effectively



**Fig. 1** The overall framework for the digitizing production in discrete manufacturing

display these data and then deliver its in-depth value to users is an very important mission of data visualization.

How to quickly dive below the surface of the information contained in data is a key problem to be solved in the era of big data, so the importance of data visualization in this process is undoubtedly.

### 3 Data Visualization

Data visualization is the use of computer graphics and image processing techniques to transform raw data sets and convert them into understandable and interactive graphical processes. Data visualization is an emerging discipline that can provide people with the opportunities to observe data and explore the value behind the data [3]. Data visualization transforms boring words or numbers into more understandable graphical and colorful data visualization graphics, helping users to understand the

value contained in the data. It can be divided into 3 processes: (1) data exchange: the original data is converted into a data structure table, and the data will be dried and cleaned before the conversion, thereby improving the validity of the data, while the traditional data model is easily affected by the error of the original data. (2) visualization mapping: this process is the key to the visualization process. In this process, the data table maps the abstract values, geographic coordinates, data relationships and other information in the data into visualization elements through specific mapping rules, which will be more easily accepted and understood by users and help them discover the laws behind the data. (3) view transformation: transform visual elements into views of terminal devices, thereby users can access information and perceive data laws through the web interactive interface.

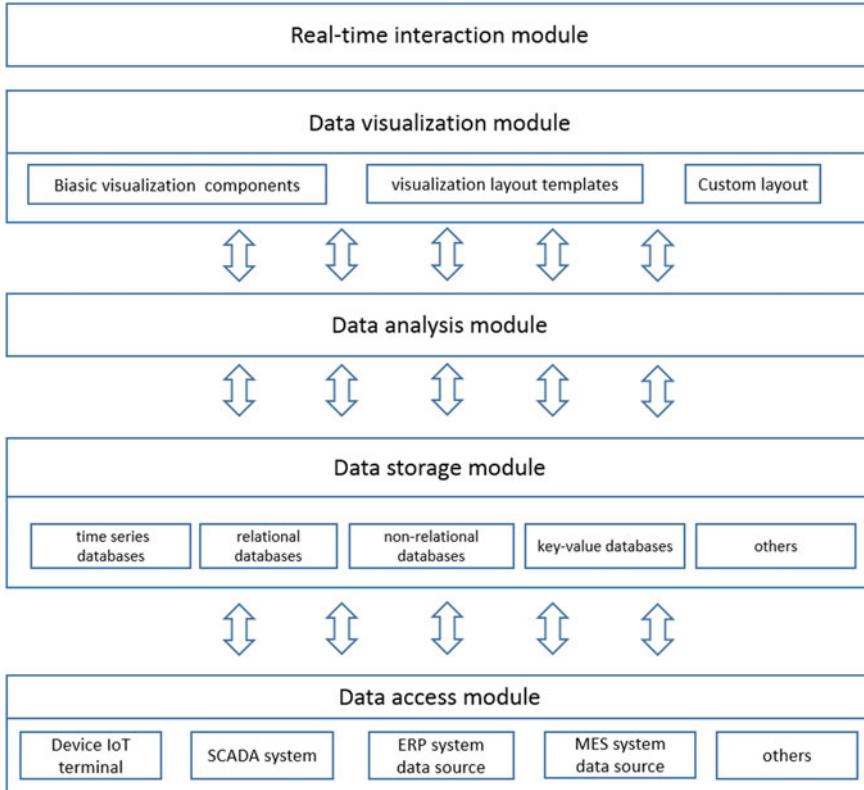
There are already a number of general-purpose and commercial data visualization tools in the industry, such as Baidu's Sugar, FanRuan's FineBI, Alibaba's AntV, Tableau's Tableau, etc.

These commercial tools have been popular in the market because of their great graphical presentation and compatibility with different types of data, whereas they do not make targeted optimization solutions for discrete manufacturing currently. Therefore, this paper designs a data visualization system based on B/S architecture especially for digital workshop management of discrete manufacturing, adding some functional modules applicable to discrete manufacturing such as real-time monitoring of equipment and dynamic supervision of production tasks on the basis of general-purpose visualization system.

## 4 The Design of the Data Visualization System

This data visualization system consists of four major functional modules: data access and storage module, data analysis module, data visualization module and real-time interaction module. The system architecture is shown in Fig. 2.

1. Data access and storage module. The system can directly access the device side IOT terminal, SCADA system, ERP system, MES system and other information system data. The system designed a set of data middleware, which can support socket protocol direct connection, Sql file import, excel format import, and intermediate table reads and writes, API gateway access and other forms. After data conversion, different types of connected data are stored in different types of databases based on the characteristics of the data types, including time series databases, relational databases, non-relational databases, key-value databases and other basic databases.
2. Data analysis module. The system is designed with a set of data analysis modules that can be dynamically loaded into the algorithm models and all algorithm models are designed with a unified data interface, that is, the input content and output results are data tables in the database, where the input table is a temporary table that can be deleted after the analysis is completed, and the output table is

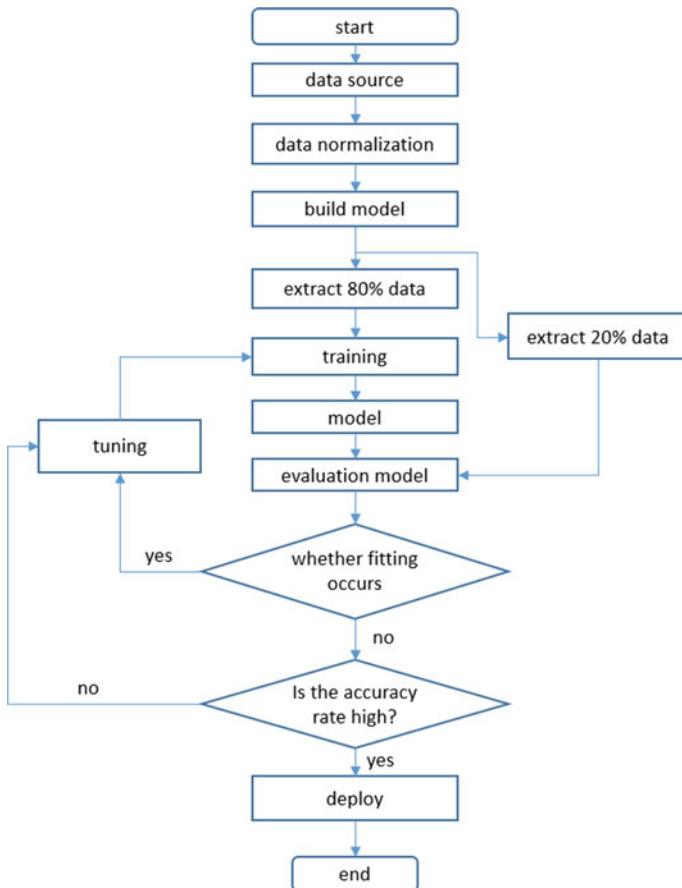


**Fig. 2** The data visualization system architecture

a permanent table that will not be deleted. The system will have several classic algorithms built-in for users to call, such as regression models, neural network learning modules, etc. Among them, the built-in regression model algorithm training diagram is shown in Fig. 3.

The built-in algorithm can support cloud upgrades. Users can customize the relevant algorithms and introduce them into the system as dynamic codes for dynamic invocation.

3. Data visualization module. The system will encapsulate commonly used visualization components, including but not limited to dynamic tables, bar charts, pie charts, line charts, maps, card flops, gauges as basic components for users to call, and a number of visualization layout templates will be built in the system. Users can also customize the design of the visualization view and embed it into the module for use. The design of the visualization module will follow the web framework principle of model-template-view, where the model connects to the database, the template handles the business logic, and the view displays the data



**Fig. 3** Regression model algorithm training diagram

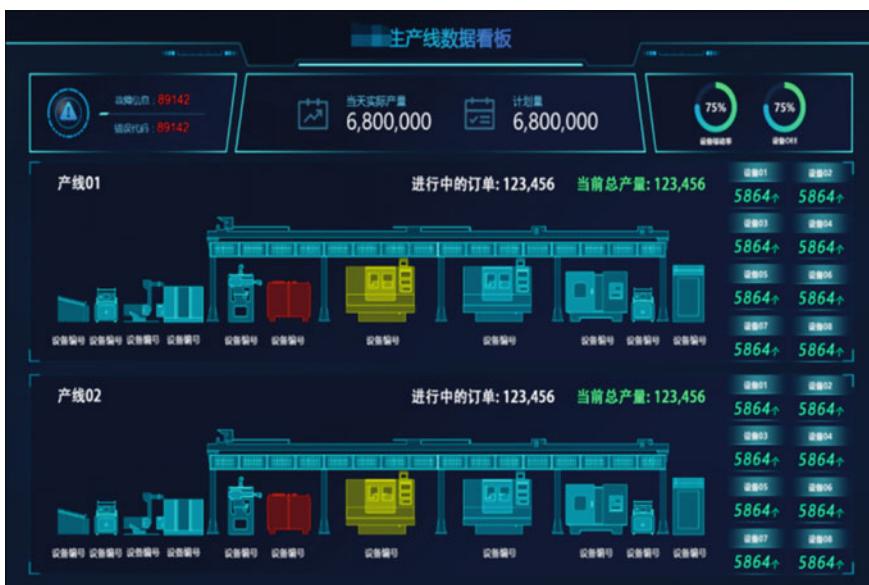
effect. In terms of human-computer interaction, the habits of discrete manufacturing industry practitioners will be fully considered to reduce the complexity of operation as much as possible, so that the data display will be shown to users in the most intuitive and convenient form. The module should be compatible with PC terminal and mobile terminal.

4. Real-time interaction module. Since in the process of workshop production, users often give feedback, modify and adjust visualization data according to the actual progress of production, the operation may be in the form of buttons, embedded panels, web pages, mobile terminals, etc., thereby the real-time interaction module is the last piece of the business closed-loop version of the entire

visualization system. In order to meet a variety of business scenarios, the real-time interaction module should have the compatibility of underlying communication and the real-time data feedback, so the core data of the real-time interaction module should be stored in memory and use a memory database to improve feedback efficiency.

5. Features of the system. One of the features is the real-time monitoring of equipment: the system distinguishes the real-time production status of equipment by different colors in the form of graphics. What is more, it can make logical grouping and logical group line for the equipment and inform users about the status of the equipment they manage at a glance, thereby they can have comprehensive awareness of the equipment information quickly. The design sketch is as in Figs. 4 and 5.

Another feature is the dynamic monitoring of production tasks. Real-time monitoring of the order tasks being produced, in addition to displaying the basic information of the order. It also dynamically calculates and displays the estimated delivery time, warning status, equipment capacity consolidation calculation, etc., assisting front-line production personnel and management personnel to grasp the production situation in a timely manner (Fig. 6).



**Fig. 4** Production line Kanban demo

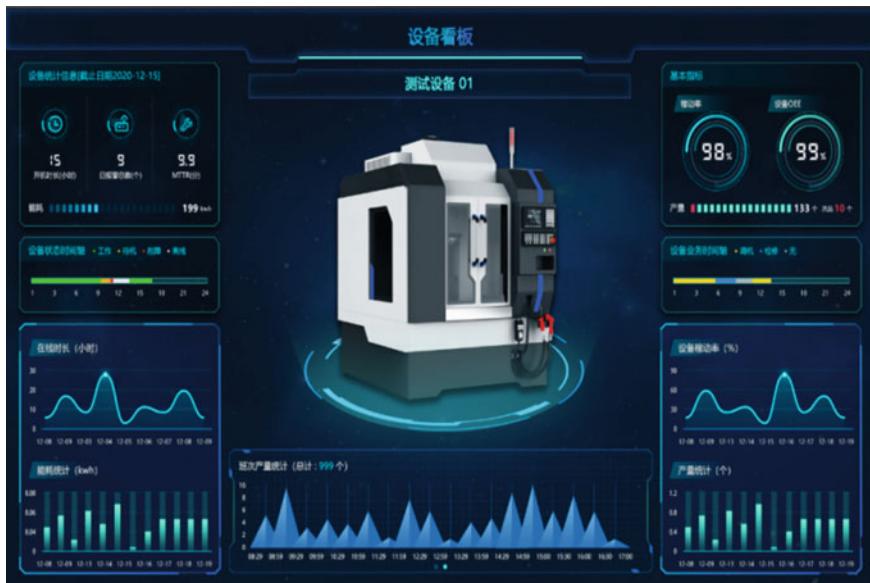


Fig. 5 Equipment Kanban demo



Fig. 6 Production tasks Kanban demo

## 5 Conclusion

The author has been engaged in digital transformation services for discrete manufacturing workshops for a long time, and has implemented and deployed data visualization systems for more than 30 local discrete manufacturing companies. Based on years of practical engineering experience along with professional knowledge, this set of visualization system for the digitalization of discrete manufacturing production has been designed, and some functions have been realized. Currently, the system has been deployed to several customer project sites and has received relatively positive user feedback. In the future, we will continue to improve the design and implementation of the system, and strive to bring value to more discrete manufacturing data visualization scenarios.

## References

1. Hu, S.J.: Evolving paradigms of manufacturing: from mass production to mass customization and personalization, pp. 3–8 (2013)
2. Lee, J., Bagheri, B., Kao, H.A.: A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manuf. Lett.* **3**, 18–23 (2015)
3. Bostock, M., Ogievetsky, V., Heer, J.: D<sup>3</sup> data-driven documents. *IEEE Trans. Visual Comput. Graphics* **17**(12), 230–2301 (2011)

# Research and Simulation of Image Specific Region Recognition Technology



Nan Li , Chang Jiang Feng , and Bin Lang

**Abstract** The development of computer technology has promoted the wide application of computer vision technology, especially in the complex background environment, accurate and fast recognition of specific areas of objects has become a hot spot of image recognition technology based on computer technology. The traditional recognition methods usually use comparison method, that is, to extract image contour by edge detection and compare with given model to complete the recognition. The efficiency of this method is low, and it is difficult to achieve efficient and fast recognition. Gabor filter of neural network is chosen as the recognition tool. For the main shortcomings of its slow recognition speed and low efficiency of extracting feature points, this paper improves the recognition method by simplifying the system model and improving the efficiency of feature point extraction, it can complete the task of real-time recognition of specific recognition targets.

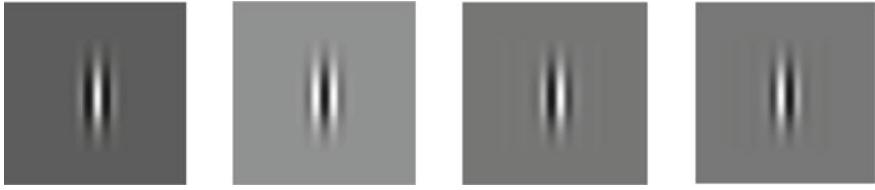
**Keywords** Gabor filter · Image recognition

## 1 Gabor Function [1]

Gabor function was first proposed by Gabor in 1946. With the improvement of research, especially in the process of studying the working mechanism of human visual system, people are more and more interested in Gabor function, This is because Gabor function can better abstract the working mechanism of visual nerve cells, and.

---

N. Li · C. J. Feng · B. Lang  
University of Army Engineering, Shijiazhuang, China  
e-mail: [Linan9585@163.com](mailto:Linan9585@163.com)



**Fig. 1** Gabor wavelet images with phase value of  $0^\circ$ ,  $180^\circ$ ,  $-90^\circ$  and  $90^\circ$ , from left to right, respectively

Gabor function can best take into account the signal resolution in time domain and frequency domain.

Definition of Gabor kernel function is showed as follows:

$$g(\lambda, \phi, \theta_k, \sigma, \gamma, x, y) = e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)} \cos\left(2\pi \frac{x'}{\lambda} + \phi\right) \quad (1)$$

where

$$\begin{cases} x' = \cos \theta_k + y \sin \theta_k \\ y' = -x \sin \theta_k + y \cos \theta_k \end{cases}$$

$k$  is the quantity of filter directions.  $\sigma$  is the standard deviation of Gaussian envelope in the  $x$  and  $y$  directions.  $\lambda$  and  $\theta_k$  are the wavelength and direction of sine wave respectively. The effective range of  $\theta_k$  is  $0^\circ$ – $360^\circ$ . Gabor wavelets from  $0^\circ$  to  $180^\circ$  are usually selected, because the wavelets separated by  $180^\circ$  are in the same straight line direction. Such as the  $\theta_k = 45^\circ$  and  $\theta_k = 225^\circ$  are the angles in a straight line. Figure 1 shows the Gabor wavelet images under different phases.

Gabor function is a Gaussian function which is modulated by complex sine function. The aspect ratio of Gaussian envelope is 1. Different from the Gaussian, the Gabor function has a balance for expansion in space with  $\sigma_x = \sigma_y$ . Gabor filter is usually described by spatial frequency bandwidth and directional bandwidth. Both the frequency bandwidth  $B$  (unit: octave) and directional bandwidth  $\Omega$  (unit: radian) of Gabor filter are defined by the range of half peak value.

$$B = \log_2[(\pi F \sigma \lambda + \alpha)/(\pi F \sigma \lambda - \alpha)] \quad (2)$$

$$\Omega = 2 \tan^{-1}(\alpha/(\pi F \sigma)) \quad (3)$$

where  $\alpha = \sqrt{\ln 2/2}$ . The Gabor filter can produce arbitrary center frequency, bandwidth and direction characteristics by changing the parameters of  $B$ ,  $\Omega$ ,  $F$  and  $\theta$ . In this way, any region with ellipse shape can be covered in frequency domain, so as to meet the requirements of extracting different texture features of the target in the image according to the requirements.

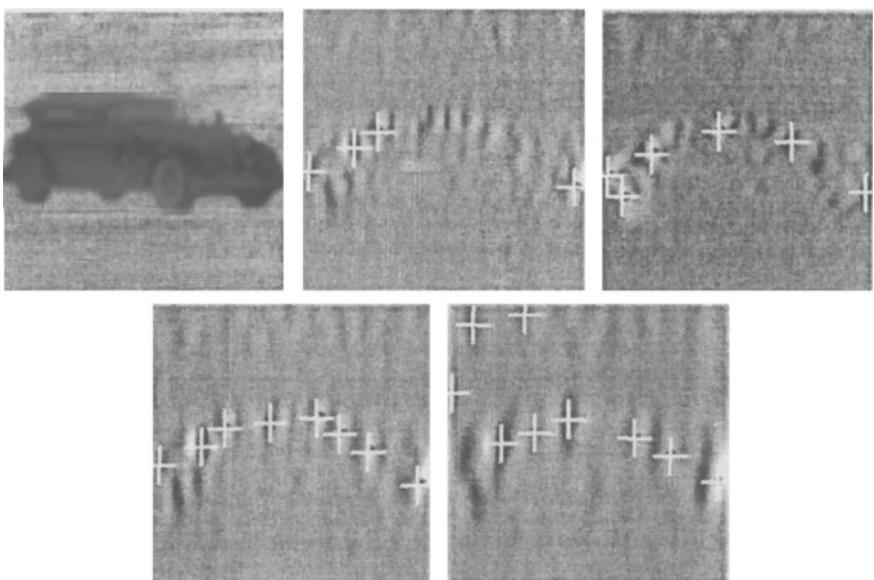
## 2 Improved Gabor Function

Gabor function can select feature samples from any number of scales. The number of feature points extracted from different scales is different (see Fig. 2). The increase of the number of feature points will greatly improve the accuracy of recognition, But it does not mean the more the better. On the contrary, too many feature points will lead to slow matching speed and longer processing time. Reducing the number of feature points can improve the real-time performance of recognition, but it is easy to reduce the recognition accuracy. In order to ensure the real-time and accuracy of recognition. How to select the appropriate scale to control the number of feature points and find the “good feature points” which can accurately reflect the position characteristics of combat vehicles has become a key problem to be solved.

A good feature point detection method has the following characteristics:

- (1) Feature point detection should be repeatable. That means the image can be reused as the feature points after different transformations (such as rotation, scaling, etc.).
- (2) The feature points can be detected in the spatial position.
- (3) The detection algorithm is simple.

The spatial and frequency characteristics of Gabor filter and Gaussian filter are very similar after analyzing the characteristics of the kernel function. The Gaussian filter is a linear smoothing filter which chooses the weighted value according to the shape of Gaussian function. The expression of two-dimensional Gaussian kernel



**Fig. 2** Feature point extraction of different resolutions

function is as follows:

$$g(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (4)$$

where,  $\sigma$  is the filtering scale,  $x$  and  $y$  represent the coordinates of pixels in the image.

The computational complexity of two-dimensional filtering increases with the square of the width of the filtering template, while Gaussian filtering is separable, that is, the convolution of Gaussian function can be divided into two steps in large-scale Gaussian filtering. In this case. The feature point extraction efficiency will be greatly improved under the same scale. In this paper, the Gabor wavelet feature point extraction method is improved based on the property of Gaussian function. By analyzing the principle of Gaussian filtering, The Harris feature point detection algorithm is selected to improve the feature point extraction method of Gabor function.

## 2.1 Harris Feature Point Detection Algorithm

Harris feature point detection algorithm [2] was proposed by Chris Harris and Mike Stephens in 1988. It is a direct method to analyze the local gray value of image. The basic principle of the algorithm is that if the brightness change of the neighborhood where a point is located is less than the preset threshold after a small distance translation, it can be considered that the point is in a “flat area” with uniform brightness. Otherwise, the point is judged as a feature point.

There are three main reasons for choosing multi-scale Harris feature point detection algorithm to improve Gabor filter. First, it uses Gaussian function as detection window, which is consistent with Gabor kernel function. Second, It has a good effect on the repeatability of the target feature points when the spatial distribution parameter  $\phi = 45^\circ \pm (n\pi/2)$ . Last, it can better control the amount of computation of the processed image for using image second-order matrix. The Harris feature point detection algorithm has the advantages of rotation, scale change, illumination change and noise invariance. In this algorithm, a local detection window is used to move slightly along all directions in the image, and the average energy change value of the window is compared with the set threshold value. If the energy change value exceeds the set threshold value, the center pixel of the window is extracted as the feature point. The Harris feature point detection algorithm calculates based on the local auto-correlation function of the signal. It moves the image window  $w$  (usually rectangular region) to any direction by a small displacement  $(x, y)$ , and the gray level change value can be defined as:

$$Gray_{x,y} = \sum_{u,v} w_{u,v} [I_{x+u,y+v} - I_{u,v}] = \sum_{u,v} w_{u,v} \left[ u \frac{\partial I}{\partial x} + v \frac{\partial I}{\partial y} + O(u^2, v^2) \right]^2 \quad (5)$$

where:  $G_{x,y}$  is the measurement of gray level change in the window  $w_{u,v} = e^{(x^2+y^2)/\sigma^2}$ .  $I$  is the image gray function. The transformation of  $G_{x,y}$  into quadratic form is as follows:

$$Gray_{x,y} = [u \ v] M \begin{bmatrix} u \\ v \end{bmatrix} \quad (6)$$

where

$$M = \sum w_{x,y} \begin{bmatrix} I_x I_x & I_x I_y \\ I_x I_y & I_y I_y \end{bmatrix}$$

By diagonalization, the results are obtained:

$$Gray_{x,y} = R^{-1} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} R \quad (7)$$

where  $R$  is the rotation factor, and its eigenvalues  $\lambda_1$  and  $\lambda_2$  reflect the image surface curvature in the two principal axis directions. In order to avoid finding the eigenvalues of matrix  $M$ ,  $tTr(M)$  and  $Det(M)$  can be used instead of finding  $\lambda_1$  and  $\lambda_2$ . if we have the following assumption:

$$M = \sum w_{x,y} \begin{bmatrix} I_x I_x & I_x I_y \\ I_x I_y & I_y I_y \end{bmatrix} = \begin{bmatrix} A & C \\ C & B \end{bmatrix} \quad (8)$$

Then the trace and determinant values of matrix  $M(x,y)$  are:

$$Tr(M) = \lambda_1 + \lambda_2 = A + B \quad Det(M) = \lambda_1 \lambda_2 = AB - C^2 \quad (9)$$

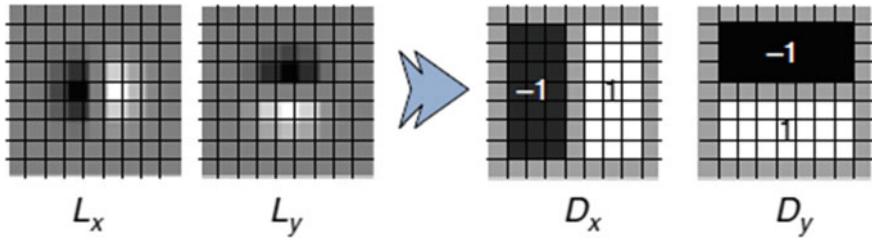
Harris' feature points response function expression is obtained:

$$R(x, y) = Det(M) - k(Tr(M))^2 \quad (10)$$

where  $k$  is a parameter greater than zero.

In practical application, the pixel is the feature point when the R value of the target pixel is greater than the given threshold.

The Harris feature point detection algorithm detects the sharp changes in the brightness of the two-dimensional image points as the feature points, these feature points have visual information, can retain the important feature information of the image, and the amount of data is small, which is conducive to improve the computational efficiency. Finding derivative in  $a$  direction for the first order Gaussian scale space of image, combining with the improved detection operator  $L_a(x)$  which is obtained by Harris algorithm, the expression can be defined as:



**Fig. 3** Approximation of wavelet response functions for  $L_x$  and  $D_x$  and  $D_y$

$$L_a(x, y, \sigma) = \frac{\partial}{\partial a} I_s(x) = \frac{\partial}{\partial a} (G(x) * I(x, y)) = (\frac{\partial}{\partial a} G(x)) * I(x, y) \quad (11)$$

$G(x)$  is the Gaussian kernel function.  $I(x, y)$  is the original image. According to formula (11), the first derivative of Gaussian function can be approximated as a set of new wavelet response functions (as shown in Fig. 3). In order to keep the scale consistent, the scale of each wavelet corresponds to the scale of Hessian detection algorithm.

Harris operator has the characteristics of simple calculation, uniform and reasonable distribution of extracted feature points, quantitative and etc. The feature points extracted through Gaussian scale space can be extracted quickly and more stable.

After the above processing, the result  $r$  of the maximum operation is written as follows: [3–6]

$$r = \max_{x_j \in S} \{ \det(H_{approx}^0(x, y, s)), \det(H_{approx}^{\pi/4}(x, y, s)), C_{approx}(x, y, s) \} \quad (12)$$

Here  $S$  is the sub window composed of pixels  $\Delta x$ ,  $\Delta y$ ,  $\Delta s$ ,  $s$  is the scale factor, and  $C_{approx}$  is the Harris matrix. In this way, the improvement of feature extraction is completed.

### 3 Feature Matching Process of Specific Target

Based on the structure of Gabor filter and its filtering effect, the number of feature points extracted by Gabor filter is proportional to the resolution, and the cost is to increase the operation time. Due to the large amount of information in the feature database of the target to be identified, the feature points in the sample database can be artificially divided into low resolution and high resolution by using Gabor filter to support multi-resolution. Under the condition of low resolution, the recognition target is classified with less number of feature points (with less cost). After the information classification of feature points is completed, the feature points are extracted from the region of interest under the condition of high resolution, and the feature points are

matched, and finally the target to be recognized is determined. This can reduce the operation time and improve the efficiency of the algorithm.

### ***3.1 The Matching Is Based On The Sum Of Energy Values In The Neighborhood Of Feature Points Under Low Resolution***

Low resolution is mainly used to recognize the contour of a specific target, not the specific feature details. Gabor filter classifies the targets according to the different energy values in the target feature points of the image. The basis of classification is to find the energy value of the feature points of the target samples to be identified in the database. Although the energy value of each feature point is different, the energy value of a single feature point can only reflect the image information from one point. In order to accurately and comprehensively express the energy value information of the image, the sum of the energy values of a region centered on the point will be used as the energy information.

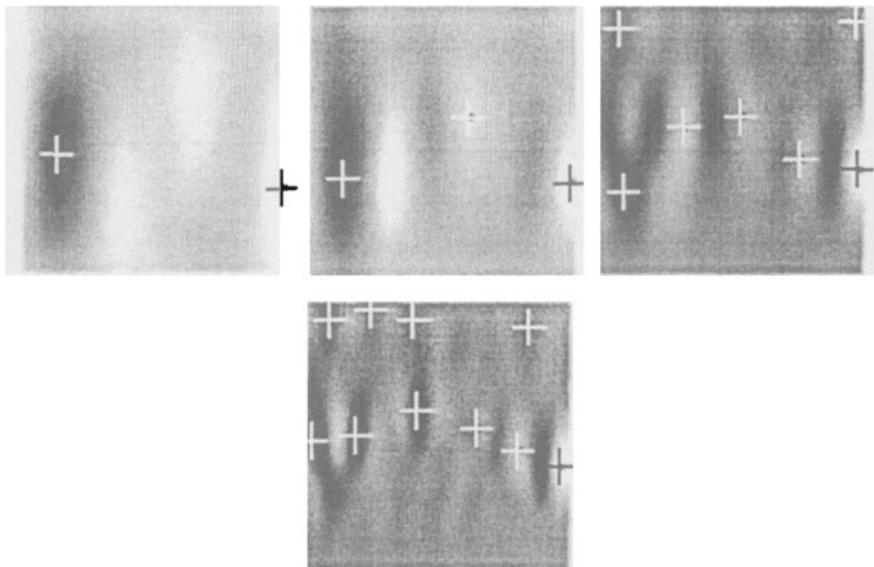
Suppose that the sum of energy values in the neighborhood of  $3 \times 3$  around each feature point is  $E_i$  at different resolutions:

$$E_i = \sum_{m=x-1}^{x+1} \sum_{n=y-1}^{y+1} E_{i(m,n)} \quad (13)$$

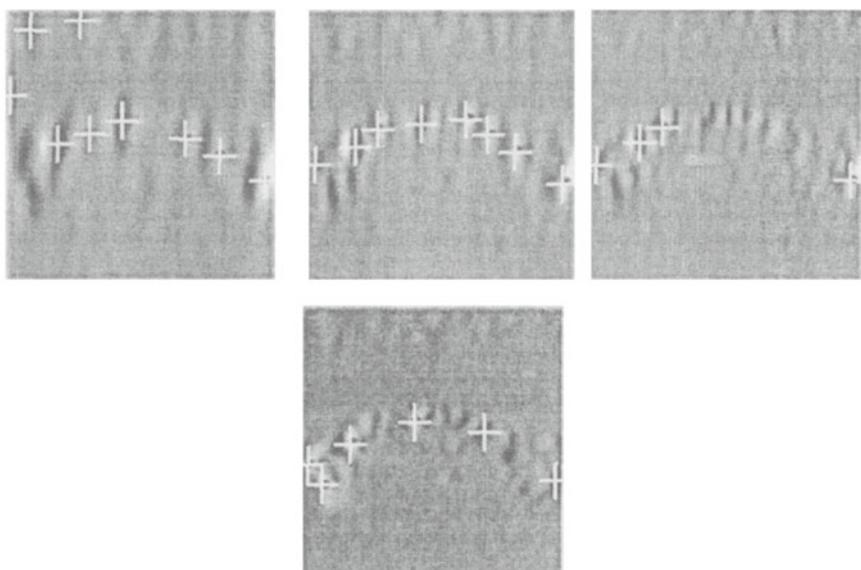
where  $(x, y)$  is the coordinates of the feature points,  $E_{i(m,n)}$  is the energy value, and  $(m, n)$  is the coordinates of the neighborhood points in the  $3 \times 3$  neighborhood. According to the above-mentioned feature point extraction method, the following samples are used to represent the sum of  $E$  values of each feature point extracted in each scale in the  $0^\circ$  direction of the filter and its surrounding  $3 \times 3$  neighborhood. For example in Fig. 4, take the tanks as specific targets. It is assumed that scale 1, scale 2, scale 3 and scale 4 are scales in low resolution, and scale 5, scale 6, scale 7 and scale 8 are scales in high resolution (Table 1).

According to the extracted feature points and the corresponding energy values of the feature points, there is a large difference in the energy between the feature points of different samples in each scale, so the difference can be used to match the feature points of the target in the image.

The principle of the feature point matching algorithm is that in low resolution, the total energy in the neighborhood of the feature points extracted from the target to be identified are arranged in the order from high to low, and the feature points matched with their energy sum are searched one by one in different directions and the same scale of the samples in the sample database. In the experiment, the points in the neighborhood of the feature points in the target sample library and the target to be identified whose energy sum error is within the range of  $\pm 10\%$  are counted



(a) Feature points extracted from tank in low resolution



(b) Feature points extracted from tank in high resolution

**Fig. 4** Feature points extracted from tank in different resolutions

**Table 1** Feature points extracted from tank in different resolutions

Scale										
Scale 1	2701.16	1026.02								
Scale 2	1669.76	711.88	474.89							
Scale 3	883.85	557.49	764.36	342.22	284.45	174.32	205.77			
Scale 4	571.84	816.3	816.3	351.02	416.23	102.19	380.29	110.82	233.54	102.63
Scale 5	91.73	97.73	433.07	273.31	480.89	215.42	160.73	129.38	73.18	
Scale 6	106.5	103.93	231.36	231.36	67.47	137.917	207.84	105.16		
Scale 7	80.78	124.43	73.84	58.13						
Scale 8	144.63	129.47	86.52	73.35	100.52	66.16				

as matched points [7–10]. According to the rules of feature points extraction, the matching rate of feature points is obtained at low resolution:

$$R = \frac{\text{The number of matched feature points}}{\min(\text{Sample feature points}, \text{Feature points with recognition target})} \times 100\% \quad (14)$$

The target can be further matched at high resolution if the R value is more than 60%. Generally, the results obtained by the matching algorithm in low resolution are not unique, that means there will be many samples' matching rate of the target to be identified greater than the setting R value in the experiment, which demonstrates that the matching can only be rough in low resolution. In this case, the matching values are low for the different types of samples or samples with large differences. On the contrary, samples with small differences need to continue to match at high resolution. In low resolution, many useless samples in the sample database can be eliminated, which can greatly reduce the amount of computation in high resolution and it plays a very important role in improving the matching speed. The matching results in low resolution are shown in Table 2.

### 3.2 Secondary Matching in High Resolution

After preliminary matching in low resolution, The Precise matching is needed in high resolution. Since the range of samples that need fine matching has been determined in

**Table 2** Match results of target and samples in low resolution

Samples in the database	Filtering direction (°)	Low resolution matching results (%)
	0	46
	45	52
	90	53
	135	47

the sample database during the coarse matching, we can continue to use the feature point neighborhood energy value to further carry out the correlation matching of feature points.

At high resolution, the feature points extracted from the target to be identified and the feature points in the target database are matched by the energy value correlation. The matching principle is that the difference of the total energy in the neighborhood of the feature points in the target sample library and the target to be identified within  $\pm 10\%$  are regarded as matching points [11–13]. The high-resolution matching not only focuses on the contour but also focus on the more detailed identification of combat vehicle contour, including gun barrel, crawler, etc. It should be pointed out that the feature points extracted by Gabor filter are different at different resolutions, that is, the feature points extracted at high resolution do not necessarily include the feature points extracted at low resolution, so it is still necessary to calculate the feature points at high resolution according to the corresponding scale. If there are more than 85% matching domain values  $r$  in the sample database and the target to be identified, the sample with the largest matching domain value  $R$  is taken as the recognition result of the system.

Table 3 shows the matching results of the target to be identified after preliminary selection at low resolution and recognition at high resolution.

In order to verify the stability of the feature extraction algorithm, 100 samples in the database are directly taken as the samples to be identified. Through the extraction of the target feature points to be identified and the matching experiment of the feature

**Table 3** Match results of target sample database samples in high resolution

Samples in the database	Filtering direction (°)	High resolution matching results (%)
	0	73
	45	77
	90	79
	135	74

**Table 4** Recognition success rate after twice recognition

Samples in the database	Filtering direction (°)	Matching times	High resolution matching results (%)
	0	100	99
	45	100	97
	90	100	98
	135	100	96

points, the recognition effect in the simplest case is verified. The final recognition results are shown in Table 4. The experimental results show that the recognition result is higher than 95%, which can be considered that the sample is directly taken as the target to be identified. The result of target recognition is accurate and reliable.

## 4 Simulation Results and Analysis

In the simulation experiment, 20 images are compared with the target recognition obtained by the traditional edge operator method and the method in this paper, and the results are shown in Table 5. The recognition rate refers to the rate of correct recognition of the target. Average computing time refers to the computing time of an image. Simulation running environment: core i5, 2G memory, 512 M video memory, windows7, matlab 2008a.

It can be seen from Table 4 that:(1) the recognition rate of traditional methods is relatively low due to the overlapping of boundary and background (2) If the edge extraction is done many times and the gradient threshold is set to complete the contour extraction, the recognition rate may be improved, but at the same time, the recognition time will increase, which will affect the subsequent process (3) Through the observation experiment, it is found that under the condition of complex background, compared with the traditional algorithm, the recognition rate of this algorithm is

**Table 5** Comparison of experimental results with different methods

Method	Recognition rate (%)	Average usage time (s)
Traditional method	55	2.3
Twice edge detection	85	3.1
Method in this paper	90	1.1

improved by 35 percentage points, and the detection accuracy is improved significantly, which can meet the requirements of accurate target recognition at higher speed. This algorithm can be combined with the traditional comparison method, which is more suitable for the requirements of short-range high-speed real-time processing, and can complete the rapid analysis and judgment of the target.

## 5 Conclusion

The research work of this paper is to find an efficient feature extraction algorithm, which not only needs to accurately identify the specified target, but also needs to meet the real-time requirements. According to the classification characteristics of Gabor filter, it is improved to simplify the calculation model, improve the efficiency of feature point extraction, and make it better meet the requirements of real-time processing. The correctness and effectiveness of the improved method are verified by simulation and experiment.

## References

1. Ingo, J.W., Wurtz, R.P.: Image reconstruction from gabor magnitudes. In: BMCV2002, LNCS, vol. 2525, pp. 117–126 (2002)
2. Jian, G., Xinhua, H., Gang, P.: A quick feature detecting method applied in robot vision. In: Proceedings of the IEEE International Conference on Mechatronics and Automation **12**, 1605–1610 (2007)
3. Zhi, W., Saixian, H.: An adaptive edge-detection method based on Canny algorithm. *J. Image Graph.* **9**(8), 957–962 (2004)
4. Hainaut, J.L.: Research in database engineering at the University of Namur. *SIGMOD Rec.* **32**, 142–145 (2003)
5. Hongsheng, X., Liangu, W., Changsong, L.: Application study of an improved method of images texture feature extraction based on gabor wavelet transform. *J. Yunnan Univ. Nat. Nat. Sci. Ed.* **18**(4), 287–291 (2009)
6. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**(8), (1996)
7. Roelfsema, P.R.: Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* **29**, 203–227 (2006)
8. Li, Z.: Remote sensing image feature and classification based on gabor function. *Geomat. Spat. Inf. Technol.* **35**(4), 123–124, 127 (2012)
9. Jeong, S., Howat, I.M., Ahn, Y.: Improved multiple matching method for observing glacier motion with repeat image feature tracking. *IEEE Trans. Geosci. Rem. Sens.* **55**(4) (2017)
10. Yuanxiu, X., Dengyi, Z., Jianhui, Z., et al.: Robust fast corner detector based on filled circle and outer ring mask. *IET Image Process.* **10**(4) (2016)
11. Qianfei, Z., Jinghong, L.: Automatic orthorectification and mosaicking of oblique images from a zoom lens aerial camera. *Opt. Eng.* **54**(1) (2015)
12. Yang, G., Zaifeng, S., Pang, K., Qingjie, C., Suying, Y.: Fast image registration based on extracting key feature points. *Acta Sci. Nat. Univ. Nankaiensis Nat. Sci. Ed.* **53**(2), 56–61 (2020)
13. Chatterji, B.N., Biswas, P.K., Manthalkar, R.: Rotation invariant texture classification using even symmetric gabor filters. *Pattern Recogn. Lett.* **24**(12) (2003)

# Guided Filter in Least Squares to Remove Non-uniform Strong Noise of Underwater Target Image



Guang Liu , Shikang Wu , Yu Shi , and Xia Hua

**Abstract** Underwater target detection will be interfered by underwater non-uniform strong noise (organic matter, suspended particles, etc.), in order to solve this problem, a new denoising method (DF-LS) is proposed based on the method of embedding bilateral filter (BLF-LS) in least square method to keep image edge smooth. Firstly, the guiding filter and the fast edge-preserving two-dimensional smoothing filter based on indicator function are introduced into the model to retain the edge of the underwater target and remove some small noise. At the same time, the guiding filter is embedded into the least square model to make the algorithm more efficient. Secondly, according to the non-uniform characteristics of noise, the multi-scale iteration of characteristic pyramid is adopted to decompose the noise of different sizes iteratively, and in the process of multi-scale decomposition, the noise of different scales is decomposed and removed step by step. Experimental results show that, compared with other classical denoising methods, the proposed algorithm can better remove the underwater non-uniform strong noise and retain the target information.

**Keywords** Non-uniform strong noise · Target image · Edge preservation · Multiscale decomposition · Steering filter

## 1 Introduction

The ocean occupies the vast majority of the earth, containing huge resources and energy. For China, which has a large population and limited land resources, abundant marine resources can alleviate many difficulties that China is facing to a certain extent. When people explore the ocean, they often take underwater photography to obtain valuable information. However, due to the complex underwater environment, the

---

G. Liu · S. Wu · Y. Shi · X. Hua ()

School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430205, China

e-mail: [21903010051@stu.wit.edu.cn](mailto:21903010051@stu.wit.edu.cn)

Hubei Key Laboratory of Optical Information and Pattern Recognition, Wuhan Institute of Technology, Wuhan 430205, China

existence of aquatic organisms, suspended solids and so on, the underwater images are subject to too much interference, resulting in non-uniform strong noise. The difficulty of underwater imaging and the complexity of the underwater environment seriously affect people's research on the ocean. Therefore, it is particularly important to remove the noise of underwater images effectively.

However, the imaging environment of underwater camera is complex, and the light propagation in the water body will be strongly attenuated. There are two main modes of attenuation: the selective absorption of light by water and the scattering of light [1]. The selective absorption of light by water can cause a considerable loss of light energy. There are two kinds of scattering of light by water, namely, the scattering of light by pure water itself and the scattering caused by substances in water. The scattering of water to light usually shows that the camera imaging will produce atomization effect, and the scattered light affects the imaging contrast, which greatly reduces the contrast. In order to obtain a clear underwater images, many methods for underwater scattering removal have been proposed [2–4]. In addition, the superposition of factors such as the flow of water, non-uniform illumination, particles and plankton in the water [5, 6] will have a great impact on the imaging of the camera, increasing the non-uniform noise of the target image and interfering with the detection and discovery of the target. At present, there are few methods to remove the non-uniform strong noise in the underwater target image, Shubin et al. [7] proposed a multi-directional morphological filtering algorithm to remove the speckle noise; He Wei used Daubechies wavelet to study the denoising of underwater images [8], but these methods are not ideal for the removal of strong noise. In order to better observe and detect the target, we need to propose a new method to remove the influence of underwater non-uniform noise on the target image.

In order to remove the noise, the underwater target is better presented in the image. In this paper, we propose an effective method to preserve the edge of the target and remove the noise: firstly, based on the image edge smoothing algorithm (BLF-LS) [9] embedded in the least square method, we propose an improved image edge smoothing algorithm (GF-LS) embedded in the least square method. In the model, the guided filter (GF) [10] is introduced as the constraint term to better retain the edge of the target in the underwater image, so that the loss of target information is less in the global processing. Secondly, according to the non-uniform characteristics of underwater noise, we iterate the feature pyramid of the DF-LS algorithm in this paper, so that the non-uniform noise in the image will be decomposed by multi-scale, and the noise of different scales will be removed by each iteration. Through the experimental test, the algorithm in this paper can remove the non-uniform noise and retain the underwater target information.

## 2 Non-uniform Noise Removal Method

### 2.1 Least Squares Model

Because our main purpose is to make the target in our underwater image appear more clearly, so this paper uses the edge-preserving smoothing method to remove the non-uniform noise of underwater interference target. Embedding bilateral filter in least squares (BLF-LS) is to use bilateral filter to constrain the global gradient feature of image and the efficiency advantage of least squares model to smooth edge. The mathematical model can be defined as (1):

$$\min_{\mu} \sum_s \left( (\mu_s - g_s)^2 + \lambda \sum_{* \in \{x, y\}} (\nabla \mu_{x,s} - (f_{BLF}(\nabla g_*))_s)^2 \right) \quad (1)$$

where the input is  $g_s$ ,  $\mu_s$  as the output,  $f_{BLF}(\nabla g_x)$  and  $f_{BLF}(\nabla g_y)$  represent the use of bilateral filter (BLF) to smooth the input image  $g_s$  gradient in x-axis and y-axis, respectively. According to Formula (1), FFTs is used to solve the problem:

$$\mu = \mathcal{F}^{-1} \left( \frac{\mathcal{F}(g) + \lambda \sum_{* \in \{x, y\}} \overline{\mathcal{F}(\nabla_*)} \cdot \mathcal{F}(f_{BLF}(\nabla g_*))}{\mathcal{F}(1) + \lambda \sum_{* \in \{x, y\}} \overline{\mathcal{F}(\nabla_*)} \cdot \mathcal{F}(\nabla_*)} \right) \quad (2)$$

where  $\mathcal{F}(\cdot)$  and  $\mathcal{F}^{-1}(\cdot)$  are fast Fourier transform (FFT) and inverse fast Fourier transform (IFFT) operators.  $\overline{\mathcal{F}(\cdot)}$  denotes the conjugate of  $\mathcal{F}(\cdot)$ . The main function of Formula (2) is to achieve image smoothing and edge preserving, which has a certain effect on noise removal. But this algorithm is global smoothing, while removing some noise, some details of the target will also be smoothed; while the underwater noise is non-uniform, the effect of this algorithm in removing strong noise is not very ideal, and through the experiment, it is found that the details of the target are smoothed. Therefore, this paper will optimize the BLF-LS algorithm model, so that the new algorithm can be applied in underwater image denoising.

### 2.2 Guided Filtering

Generally speaking, the gradient of pixels around the noise changes greatly, and the gradient around the noise is similar. The gradient step appears at the edge, and the direction of the maximum gradient is in the normal direction of the edge, and the other direction away from the normal direction gradually decreases. The general filtering cannot distinguish the noise and edge, so it is unified to deal with it, so in many cases, while filtering, the edge is also blurred. In order to remove the noise and better retain the edge of the target, we will use the guided filter.

We can think that the image is a two-dimensional function, so we can assume that the output and input of the function satisfy the following linear relationship in a two-dimensional window:

$$q_i = a_k I_i + b_k, \quad \forall i \in w_k \quad (3)$$

where  $q$  is the value of the output pixel,  $I$  is the value of the input pixel, and  $I$  and  $K$  are the pixel indexes. In this case, we can take the gradient on both sides of the above formula, and we can get it from (3):

$$\nabla q = a_k \nabla I \quad (4)$$

Here  $\nabla$  is the operator. It can be seen that the gradient of the output and the input also retains the linear relationship, so the guided filter has the characteristics of edge preservation. In order to solve the coefficients  $a_k$  and  $b_k$  in (3), it is assumed that  $p$  is the result before  $q$  filtering and the difference between  $q$  and  $p$  is minimized:

$$E(a_k, b_k) = \sum_{i \in w_k} ((a_k I_i + b_k - p_i)^2 + \epsilon a_k^2) \quad (5)$$

By optimizing (5), the linear coefficients  $a_k$  and  $b_k$  are defined as:

$$a_k = \frac{\frac{1}{|w|} \sum_{i \in w_k} I_i p_i - \mu_k \bar{p}_k}{\sigma_k^2 + \epsilon} \quad (6)$$

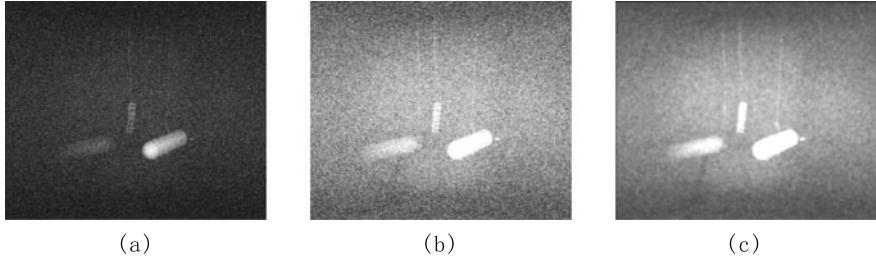
$$b_k = \bar{p}_k - a_k \mu_k \quad (7)$$

where  $\mu_k$  and  $\sigma_k^2$  are the average and variance of the pixel values of  $I$  in the window  $w_k$ , respectively,  $|w|$  is the number of pixels in the window  $w_k$ , and  $\bar{p}_k$  is the average value of the pixel values of the image  $p$  to be filtered in the window  $w_k$ .

Through the coefficients  $a_k$  and  $b_k$ , we can see that there are two inputs of the guided filter, one is the real input  $p$ , the other is the guided input  $I$ , and the output  $q$  is the product of the joint action of  $p$  and  $I$ . In this way, the filtering effect can restore the real target more effectively and reduce the loss of edge information.

## 2.3 Embedding Guided Filter Model in Least Square Method

Inspired by the edge smoothing property of BLF-LS algorithm, this paper proposes a de-noising algorithm which embeds steering filter in least square method. In order to reduce the influence on the target edge and detail information in the process of denoising, we need to smooth the noise near the edge while preserving the details on the edge and target. In this paper, the pilot filter proposed by He et al. [10] is used,



**Fig. 1** **a** The original. **b** The gradient of the original image. **c** The gradient map of the original image after EP filtering [11]. By comparing (**b**) and (**c**), it can be found that the EP filtered image as a guide image can help image restoration better

and the experimental data show that (as shown in Fig. 1), the image processed by the method proposed in reference [11] is used as the pilot map of the pilot filter, so the new mathematical model is:

$$\min_{\mu} (\mu - p)^2 + \lambda (\nabla \mu - (f_{DF}(\nabla p, \nabla f_{EP}(p))))^2 \quad (8)$$

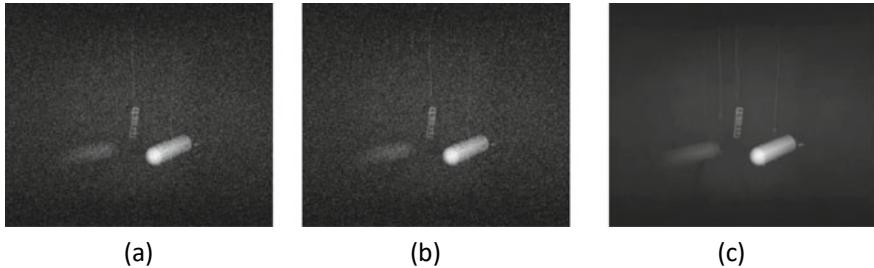
where,  $\mu$  and  $p$  are the output and input images respectively,  $\nabla$  is the gradient operator, containing the gradient in the vertical and horizontal directions,  $f_{DF}$  and  $f_{EP}$  respectively represent the use of the guide filter and exponential function-based fast edge-preserving smoothing filter in Literature [11]. According to Formula (8), it can be concluded that:

$$\mu = \mathcal{F}^{-1} \left( \frac{\mathcal{F}(p) + \lambda \overline{\mathcal{F}(\nabla)} \mathcal{F}(f_{DF}(\nabla p, \nabla f_{EP}(p)))}{\mathcal{F}(1) + \lambda \overline{\mathcal{F}(\nabla)} \cdot \mathcal{F}(\nabla)} \right) \quad (9)$$

where  $\mathcal{F}(\cdot)$  and  $\mathcal{F}^{-1}(\cdot)$  are fast Fourier transform (FFT) and inverse fast Fourier transform (IFFT) operators.  $\overline{\mathcal{F}(\cdot)}$  denotes the conjugate of  $\mathcal{F}(\cdot)$ . In the process of solving  $\mu$ , in order to get rid of the non-uniform strong noise, we use the iterative method to decompose the input  $p$  into multi-scale, and put the result of each denoising into the next solution. In this way, we can get rid of the noise better by gradually increasing the scale in the iterative process.

### 3 Experimental Results and Analysis

In order to verify the effectiveness of the algorithm, this paper first uses simulation experiments to compare the BLF-LS [9] filtering algorithm model and the denoising filtering algorithm in this paper. Objectively, the performance of the algorithm is evaluated by its peak signal-to-noise ratio (PSNR), as shown in Fig. 2. However, the real underwater images are difficult to evaluate the performance of denoising algorithms,



**Fig. 2** Comparison of denoising results of two filtering algorithms. **a** Simulation diagram. **b** The results of BLF-LS filtering algorithm are given, The PSNR was 3.85. **c** The results of the algorithm are given in this paper, The PSNR was 12.63

and standard image quality evaluation standards, such as PSNR or SSIM, cannot be used. Here, we use the classical image denoising algorithm to compare with the algorithm in this paper. These classical image denoising algorithms include gradient  $L_0$  norm minimization algorithm [12] (hereinafter referred to as  $L_0$  algorithm) and non-local image denoising algorithm (NLM) [13]. The comparison results are shown in Figs. 3 and 4.

Our experiments are carried out on a PC equipped with Intel i5-7400 CPU (3.0 GHz) and 8 GB ram, and the working system is windows 10 Education. The algorithm is implemented by MATLAB 2016a and does not need any GPU acceleration.

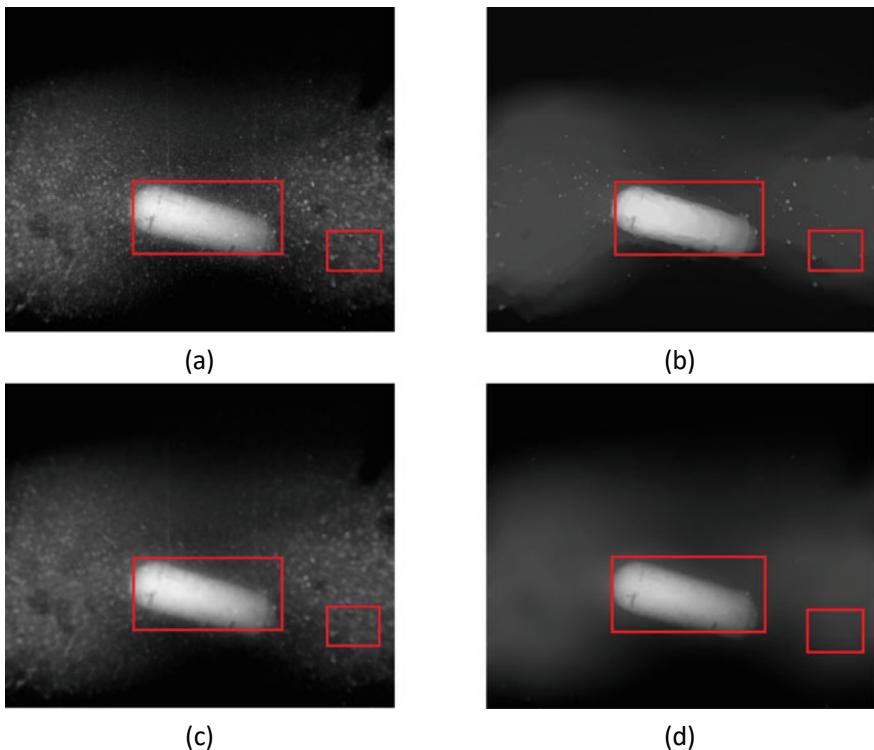
In order to quantitatively evaluate the performance of the proposed algorithm and the other two classical algorithms, this paper uses ENL as an index to measure the performance of the algorithm. ENL is used to measure the smoothness of the uniform region of the image, defined as:

$$\text{ENL} = \frac{\mu^2}{\sigma^2} \quad (10)$$

where  $\mu$  and  $\sigma$  represent pixel mean value and standard deviation of uniform region, respectively. The higher the ENL value is, the higher the smoothing efficiency of the strong noise in the uniform area is. Table 1 represents the data in Fig. 4, where ENL1 is the ENL value in the red box for the target area and ENL2 is the ENL value in the red box for the background.

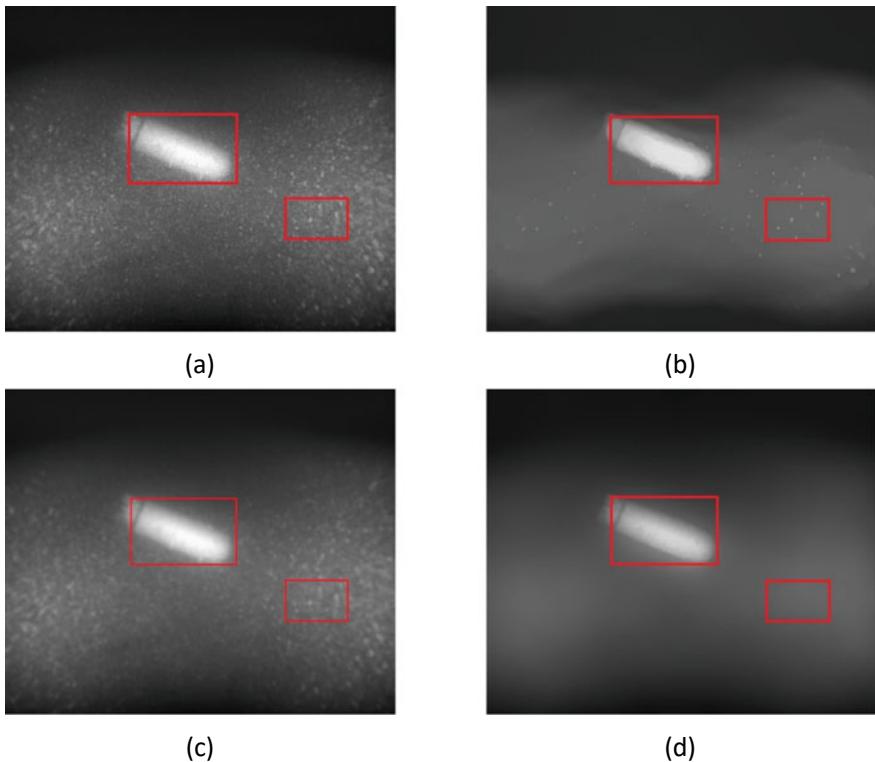
## 4 Conclusion

This paper makes some research on the non-uniform strong noise in the underwater target image. Firstly, a new optimization algorithm is proposed based on the edge preserving smoothing algorithm (BLF-LS) embedded in the least square method.



**Fig. 3** Comparison of denoising results of different algorithms. **a** Underwater target image. **b**  $L_0$  algorithm. **c** NLM filtering algorithm. **d** The results presented of the algorithm

The guided filter is introduced into the model to better protect the edge information of the target and make the image smooth and denoise globally without distortion. Secondly, aiming at the non-uniform strong noise of underwater real images, we use the method of multi-scale decomposition iteration to remove different sizes of noise on the image. Experimental results show that the algorithm has a good effect on underwater image denoising.



**Fig. 4** Comparison of denoising results of different algorithms. **a** Underwater target image. **b**  $L_0$  algorithm. **c** NLM filtering algorithm. **d** The results presented of the algorithm

**Table 1** ENL evaluation of denoising results with different algorithms

Algorithm	Original image	$L_0$	NLM	Our algorithm
ENL1 (Fig. 3)	3.97	3.62	4.01	<b>4.26</b>
ENL2 (Fig. 3)	22.72	72.81	40.21	<b>67.02</b>
ENL1 (Fig. 4)	7.37	6.96	7.77	<b>8.46</b>
ENL2 (Fig. 4)	68.71	494.56	167.67	<b>545.12</b>

Bold indicates the largest number in the experimental data, which works best

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (No.61801337).

## References

1. Sun, C., Chen, L., Gao, L., et al.: Optical properties of water and its influence on underwater imaging. *Appl. Opt.* **04**, 40–47 (2000)
2. Chiang John, Y., Chen, Y.-C.: Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* **21**(4), 69–1756 (2012)
3. Galdran, A., et al.: Automatic red-channel underwater image restoration. *J. Vis. Commun. Image Represent.* **26**, 132–145 (2015)
4. Zhiguo, F., Qiang, S., Qingqing, D., et al.: Polarization restoration of underwater target based on global parameter estimation. *Opt. Precis. Eng.* **26**(7), 1621–1632 (2018)
5. Xia, H., Chao, P., Yu, S., Jianguo, L., Hanyu, H.: Removing atmospheric turbulence effects via geometric distortion and blur representation. *IEEE Trans. Geosci. Remote Sens.* (2020). <https://doi.org/10.1109/TGRS.3043627>
6. Komuro, T., Chen, K., Enomoto, K., et al.: Tracking and removal of suspended matter from underwater video images. In: International Conference on Quality Control by Artificial Vision. International Society for Optics and Photonics (2017)
7. Shubin, Y., Dui, P., Zhiyuan, X., Yanjun, C.: Filtering algorithm of laser underwater image polluted by speckle noise. *Infrared Laser Eng.* **04**, 318–321 (2002)
8. Wei, H.: Research on underwater image denoising method based on Daubechies wavelet. Harbin Engineering University (2006)
9. Liu, W., Zhang, P., Chen, X., et al.: Embedding bilateral filter in least squares for efficient edge-preserving image smoothing. *IEEE Trans. Circ. Syst. Video Technol.* 1–1 (2018)
10. He, K., Sun, J., Tang, X.: Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1397–1409 (2013)
11. Abiko, R., Ikebara, M.: Fast edge preserving 2D smoothing filter using indicator function. In: ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2019)
12. Xu, L., Lu, C., Xu, Y., et al.: Image smoothing via L0 gradient minimization. *ACM Trans. Graph.* **30**(6) (2012)
13. Buades, A., Coll, B., Morel, J.M.: Nonlocal image and movie denoising. *Int. J. Comput. Vis.* **76**(2), 123–139 (2008)

# Optical Flow Fusion Synthesis Based on Adversarial Learning from Videos for Facial Action Unit Detection



Shuangjiang He , Huijuan Zhao , Jing Juan , Zhe Dong , and Zhi Tao

**Abstract** Human expression often happens simultaneously with head posture in real-time video, facial Action Units (AUs) is the key factor in facial expression detection. Optical flow can effectively capture weak motion displacements, that is, it can capture facial AUs caused by facial expressions. But the optical flow would produce a lot of noise, which would adverse the detection performance. To achieve a better facial AU detection performance, we propose a novel Optical Flow Synthesis Generative Adversarial Network (OFS-GAN). Firstly, we calculate the optical flow vector of the source frame and the target frame pair that randomly selected from video clips of facial expressions to promote the robustness of OFS-GAN. Secondly, in the generator of OFS-GAN, representation input into the encoder is yield by the optical flow vector concatenated with the source frame. The feature map output from the encoder is fed into the decoder to synthesize a target frame named generated target frame. In the end, through the comparison of the generated target frame and the target frame randomly selected from the target stage, OFS-GAN learns discriminative and robust facial action feature for facial AU detection following the principle of adversarial learning. Our novel OFS-GAN has been tested on DISFA+ and CK+ dataset with LOSO evaluation method. Qualitative results of our experiment demonstrate that OFS-GAN approaches or exceeds existing optical flow or deep learning algorithms.

**Keywords** Adversarial learning · Optical flow · GAN · Action unit detection · Facial expression recognsition

---

S. He · H. Zhao

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China

e-mail: [D202080766@hust.edu.cn](mailto:D202080766@hust.edu.cn)

Z. Dong · Z. Tao

Wuhan Fiberhome Integration Technologies Co., Ltd., Wuhan, China

J. Juan

Wuhan Intelligent Medica Research Institute Co., Ltd., Wuhan, China

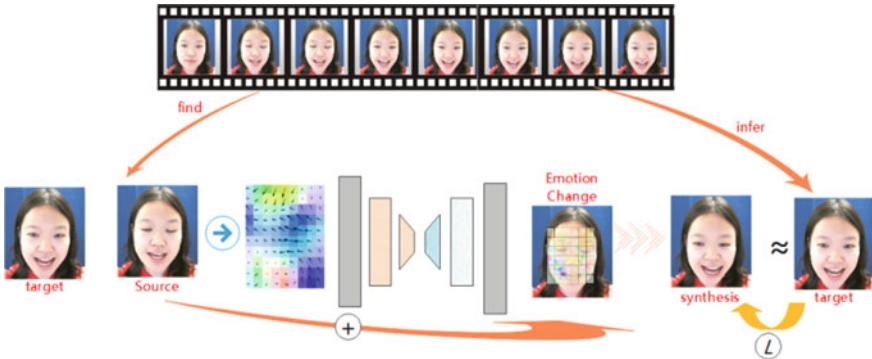
## 1 Introduction

Human face is a sophisticated representation system of human emotion, which is as complex as the human language system. Facial expressions are generated by various muscles and nerves, and occur with head postures simultaneously. For example, laughter is accompanied by a forward and backward tilt, and sadness is accompanied by drooping head in addition to drooping eyebrows and eyes. In psychological research, facial expressions and head posture movements are often associated with movements. In terms of facial movement and head posture, Ekman did a lot of work in 1970s [1]. It is proved that facial expressions and head movements jointly express human emotions, which is independent with race. Furthermore, Ekman has developed FACS (Facial Action Coding System) to define facial AUs (Action Units) to analyze the human expression change. Human beings themselves can learn expression differences through a small amount of picture training. With the development of technology, computers are used in human expressions recognition, emotional analysis, health assistance, human–computer interaction and other fields. In terms of physiological, expressions can be recognized in different human faces because they are independent with individual identity. Manual features have been used for facial AUs detection in the earlier methods [2, 3]. The booming development of deep learning facilitated automatic expression recognition and separation [4, 5], which have become major tasks in deep learning field.

Deep learning makes the end-to-end method mainstream in automatic facial expression recognition tasks. Movement information can be detected in the video, including head posture, facial actions and information. It is essential to be able to recognize subtle movements in local facial areas (e.g., eyes, eyebrows, nose, head posture) and changes in head posture, and filter out information unrelated to these movement. The main task of the separation is to remove noise and retain enough information to convey an expression.

According to the aforementioned statements, we proposed a novel Optical Flow (OF) Synthesis Generative Adversarial Network (OFS-GAN) to automatically encode AUs displacement in videos. When the system detects the expression change onset, it will be used as the source frame, and the OF calculation are performed together with the target frame to output the OF result. Firstly, we calculate the optical flow vector of the source frame and the target frame pair that randomly selected from video clips of facial expressions to promote the robustness of OFS-GAN. Secondly, in the generator of OFS-GAN, representation of the input is yield by the optical flow vector concatenated with the source frame. The feature map output from the encoder is fed into the decoder to synthesize a target frame named generated target frame. In the end, through the comparison of the generated target frame and the target frame randomly selected from the target stage, OFS-GAN learns discriminative and robust facial action feature for facial AU detection following the principle of adversarial learning. Figure 1 presents the overview of OFS-GAN.

In summary, our contributions are two folds: (1) We built a novel Optical Flow Synthesis-GAN to detect facial AUs in videos in adversarial manner. (2) OFS-GAN



**Fig. 1** Overview of the proposed OFS-GAN model

can successfully separate subtle facial expressions and subtle head posture from facial expression movement to achieve a more robust descriptor for AU detection.

## 2 Related Works

Facial expression recognition (FER) has attracted great attention of many researchers [6]. FER extracts overall features from whole face, including identity, facial action, head posture, age, race and so on [7]. The features can be divided into two types: human biological features and potential features. Traditional methods extracted human biological features, such as Histograms of Oriented Gradients (HOG) [8], optical flow information [9], Histograms of Local Binary Patterns (LBP) [10] et al., which are all manual features. However, these manual features are poorly robust and have high requirements for experimental conditions, such as specific posture and fixed illumination. With booming development of deep learning, there are many approaches in deep learning manner are proposed [11]. Deep learning methods extract more abundant facial attributes such as different poses, illumination and identity information and so on under different head postures and different expression levels. Deep learning makes FER more effective, but facial expression features often dissolve with head posture extracting expression fragments from videos or recognizing expressions in natural environment, which would adverse the performance. Obviously, FER still faces many research challenges. Hinz et al. [12] generated image features through an encoder to separate the controllable part from the redundant part. Based on these different methods, our model can separate human faces and learn the changes of expression and posture, thus improving the performance of the classifier. Generative Adversarial Network (GAN) was first proposed by Goodfellow et al. [13], which consists of a generator G and a discriminator D. In our method, by inputting the different values calculated by optical flow of two frames into the encoder, the generator uses the source frame to combine with random noise

to generate synthetic fake/negative samples. At this time, the discriminant decoding and outputs the comparison target frame to separate real/positive samples from the fake. Under the condition of satisfying MINMAX, the training model can be obtained by repeated training of G and D, the model can be trained in adversarial manner as follows:

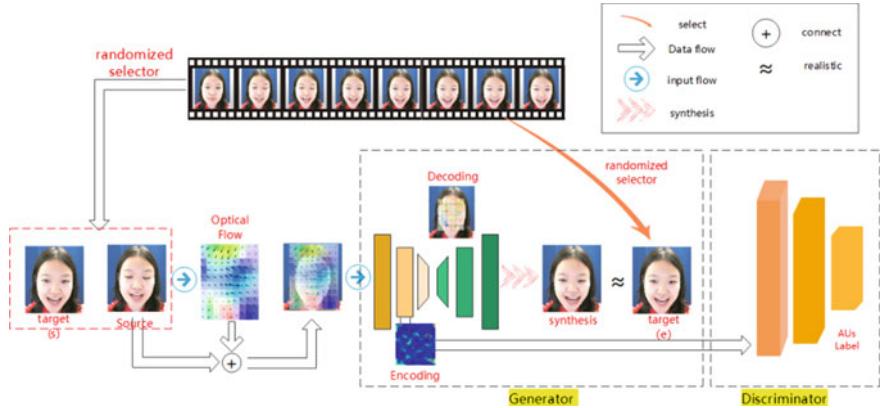
$$\begin{aligned} \min_G \max_D V(D, G) = & E_{x \sim P_{\text{data}}} [\log(D(x))] \\ & + E_{z \sim P_z} [\log(1 - D(G(z)))] \end{aligned} \quad (1)$$

where  $P_{\text{data}}$  is the data distribution of real samples and noise Z follows the distribution of  $P_z$ . GAN is currently widely used in computer vision and has a good effect in the fields of composite image and facial expression coding [14]. The Twin-Cycle coding structure can input facial pictures at disentangle moments of AUs in pairs, and realize the adversarial learning method to automatically learn the expression of facial expression and posture [15]. But there are also expressive defects in the separation of facial expressions and postures, the reason is that facial expressions are local movements within the face but head motions are relatively global movements. Following the concept of adversarial learning, our model is encouraged to disentangle AUs information use optical flow features, which yields a more discriminative expression movement for FER task in video.

### 3 Proposed Method

#### 3.1 Overview

This section describes the overview of our novel fuse end-to-end approach, which selects the source frame and target frame to compute Optical Flow to synthesis realistic face for expression changes detection. Our Optical Flow Synthesis-GAN framework is shown in Fig. 2. Firstly, it selects source frame and target frame(s) from the video. The source frame would be randomly sampled from the stage of onset and the target frame from the stage of apex similarly, and the Optical Flow is applied to both frames together to obtain the optical flow vectors. Secondly, the optical flow vectors are combined with the source frame into the module of generator which is encoded to feature map then decode to a fake realistic facial expressions output. It would be again randomly select target frame(e) to compare to the synthesis image for training the generator this moment. It is obvious that the randomized selector is helpfully obtained the distribution of facial expression of change for the generator. Meanwhile, the encoding of fake realistic facial expression would be input to the discriminator which is classify the encoded synthesis and by AUs labels. Finally, OFS-GAN is trained to obtain the model that only attention to the facial expression changes as the AUs and ignore the other facial attributes by iterator a lot of epochs.



**Fig. 2** The overall framework of our proposed optical flow synthesis-GAN. The generator as a module which would fuse the optical flow and source frame to generate a realistic facial expression, the discriminator is classifying the encoded synthesis by AUs labels

### 3.2 OFS-GAN

OFS-GAN is based on the principle of GAN, it can be divided into two parts: the generator and discriminators. In order to get the robust model by playing a minmax the two-player game, the model can be defined in the objective function as follows:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}} [\log(D(x))] + E_{s, o \sim P_{data}, z \sim P_z} [\log(1 - D(G(s, o, z)))] \quad (2)$$

where  $P_{data}$  is the data distribution of real samples,  $x$  is the picture synthesized by combining the source frame and the optical flow vectors and the noise decoding,  $S$  is the encoding of source frame, and  $O$  is the encoding of the optical flow vectors, and the noise  $Z$  follows the distribution of  $P_z$ .

In our framework, the generator determines the effect of the framework, discriminators use the encoder feature obtain the model. Given an input tuple  $(I_s, I_o, Y_{te})$ , where  $Y_{te}$  is one-hot labels of the target(e) frame,  $I_s$  is input the source frame,  $I_o$  is the optical flow vectors, we feed them into the two encoders to obtain the corresponding image representations. Concretely, we denote  $E_s$  as the face encoder and  $E_o$  as the expression encoder. Thus, the encoded representations of the two inputs can be formulated as:

$$d_s = E_s(I_s), d_o = E_o(I_o) \quad (3)$$

where  $d_s$  is the source frame representation and  $d_o$  is the representation of optical flow vectors. We then fuse these two representations through the embedding module, which can be described as:

$$d_{\text{fuse}} = \text{Emb}(\text{con}(d_s, d_o)) \quad (4)$$

where  $d_{\text{fuse}}$  denotes the fused representation,  $\text{Emb}(x)$  is the embedding module and  $\text{con}(x; y)$  indicates the operation of channel-wise concatenation. We haven't additionally introduced a noise vector into the fusing process, because the optical flow vector has some minor variations, e.g., a different angle, in generated images and makes the model more robust during training. In our design,  $d_{\text{fuse}}$  lies in a hidden space that encodes both the high-level semantics of the input source frame and optical flow vectors. Therefore, we generate the synthesis through the decoder based on  $d_{\text{fuse}}$ , which can be expressed as follows:

$$I_g = D_g(d_{\text{fuse}}) = G(I_s, I_o) \quad (5)$$

where  $D_g$  and  $G(x)$  denote the decoder and the overall generator respectively. Ideally,  $I_g$  should follow the same facial expression as target(e) frame and simultaneously with an AUs consistent with target(s) frame. By introducing the additional  $d_{\text{fuse}}$  feature, the generator gets feedback from the discriminator, which forces the generator to synthesize images that follow the data distribution of AUs dataset and helps rendering a consistent facial expression. Therefore, the loss function can be formulated as:

$$L_G = E_{(I_g, y_{te}) \sim P_a} [\log D_a(d_{\text{fuse}})] + ||I_g - Y_{te}||_1 \quad (6)$$

where  $L_G$  is the corresponding loss function of the AUs discriminator.  $P_a$  refers to the data distribution of facial expression change videos,  $y_{te}$  is the label of AUs of target frame. The AUs discriminator  $D_a \in R^{K_a}$  conducts the task of facial AUs, the  $K_a$  is the class number of AUs.

### 3.3 Randomized

In order to make the OFS-GAN more robust and improve the ability of model generalization, we randomly selected three frames from the candidate frames of stage of onset and stage of apex as the source frame and the target(s) frame and the target(e) frame. The result of random selector is shown in Fig. 3. From the first frame of stage of onset to the first frame of stage of apex as a window denote W1, and from the last frame of stage of onset to the last frame of stage of apex as a window denote W2. The length of stage of onset as denote lo and the same as the stage of apex as denote la. It is Random slide W1 to the right and also reversely slide W2 which the slide distance is equal to the small one between the lo and the la. When the sliding was achieved that the window's two endpoints are the select pair of frames. From above method the two windows could find two pairs of frames which have two source frames and two target frames, we are randomly select a source frame from two as denote source frames and select two target frames as the target(s) and target(e).



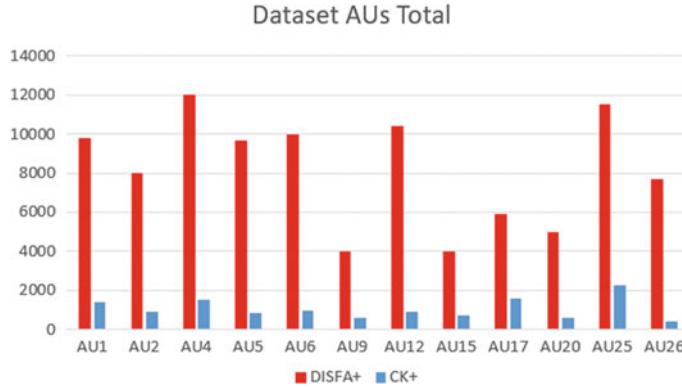
**Fig. 3** The result of Randomized selector from the video. The left image is selected to source frame, and middle one is target(s) frame, and right face is target(e) frame

## 4 Experiments

In this section, we validated the effectiveness of the proposed OFS-GAN. We have performed a series of experiments on the two conventional dynamic datasets, which show the precision improvement from the OFS-GAN compared with the Optical Flow classical feature extraction method and CNN + LSTM meth-od in the FER classification task.

### 4.1 Datasets

Our experiments adopt two major databases for facial AU detection in videos: DISFA+ [16], CK+ [17], they provide the annotation of facial expressions for AUs described FACS. DISFA+ contains 9 subjects on 12 AU and the list of 42 posed facial actions, these images sequences were digitized into  $1280 \times 720$  pixel arrays with 24-bit color values. As a complementary dataset, CK+ includes 593 FACS sequences from 123 subjects, they were digitized into either  $640 \times 490$  or  $640 \times 480$  pixel arrays with 8- bit gray-scale or 24-bit color values. However, the positive AU intensity is only in the DISFA+ database which includes 5 levels as A to E, the CK+ database's AU intensity is implicit in the emotion label. We adopt Holdout-database Evaluation (HDE) protocol. Since there are 12 AUs in the DISFA+ dataset but the CK+ dataset has 30 AUs, CK+ dataset would be clipped to 12 AUs to adapt the DISFA+ dataset. Summary of the two datasets is shown in Fig. 4. For HDE protocol, we would be used DISFA+ to train and CK+ to test, in test result to manual check AU intensity.



**Fig. 4** Both datasets are distribution of number of frames for each AU of histogram. Although the number of frames of AUs in the both databases was poles apart, the distribution proportion was similar

## 4.2 Experimental Setting

As input data Both datasets were preprocessed. The DISFA+ dataset's all of frames was copied clip of header area adjusted to  $256 \times 256$ , and the number of frames in the CK+ database was resized to  $256 \times 256$ , and the colors of all of dataset ware all grayed.

The experiments are all trained on 3080Ti GPU using the same hyper parameters: BATCH = 32, EPOCH = 100, and the optimizer is Adam with an initial learning rate of 1e-4. Leave One Subject Out (LOSO) method is used for the evaluation, using Unweighted F1-score (F1). To obtain them, we need to calculate all the True Positives (TP), False Positives (FP), and False Negatives (FN) for the k-fold verification of each AU a (12 AUs). OFS-GAN be the number of samples in class c, then the values of F1 is given by the following formulas:

$$F1_a = \frac{2 * TP_a}{2 * TP_a + FP_a + FN_a} \quad (7)$$

As a pair of inputs include two frames in OFS-GAN. Our method finds the source frame by randomly selecting a frame at the beginning stage of the expression as the source frame, and the target frame is also randomly selecting a frame at the apex stage.

The goal of our experiment is to detect AUS in video to analyze facial expression movements, requiring compared of methods capable of performing multi-tag tasks in video. Therefore, we choose the traditional Optical Flow method, CNNLSTM and AURCNN as the experimental compared of methods. To experiment the effectiveness of each attribute of the methods, the details of the compared models are summarized in Table 1.

**Table 1** Compared methods details

Model	RGB	OF	GAN	VIDEO
Optical flow RNN	X	✓	X	✓
CNN-LSTM	✓	X	X	✓
AU-RCNN-LSTM	✓	X	X	✓
OFS-GAN (ours)	✓	✓	✓	✓

\* *RGB* use RGB information, *OF* use optical flow information, *GAN* base on the model GAN in training, *VIDEO* need video context

### 4.3 Results and Analysis

The experimental results compared with the above methods are shown in Table 2 respectively. In the Optical Flow experiment, we extract the optical flow feature from the onset frame to the apex frame in the video and then input it into the RNN for AUs detection. In the CNN-LSTM and AU-RCNN experiment, each video is used to resize to same shape each expression image sequence into different numbers of frames, and then the dynamic image feature is extracted and input into CNN-LSTM and AU-RCNN for classification. In the Proposed experiment with our method, randomized select the source frame and target(s) frame fuse extract optical flow feature and combined source frame as input to OFS-GAN which use the training generator the synthesis the face then the discriminator would use the fake realistic face to classify the AUs labels. Above three methods are all extract each frame but our

**Table 2** The F1 and AR of the baseline methods and our OSF-GAN

AUs	Optical flow	CNN-LSTM	AU-RCNN-LSTM	OFS-GAN (ours)
AU1	17.5	26.3	32.1	<b>38.20</b>
AU2	18.7	23.4	25.9	<b>29.65</b>
AU4	37.2	54.2	<b>60.8</b>	57.50
AU5	33	51.15	58.05	<b>59.60</b>
AU6	28.8	48.1	55.3	<b>59.70</b>
AU9	12.7	29.9	39.8	<b>41.85</b>
AU12	37.7	69.4	67.7	<b>68.55</b>
AU15	38.1	74.75	72.55	<b>78.65</b>
AU17	35.55	62.95	<b>65.3</b>	63.13
AU20	37.85	67.15	69.1	<b>74.13</b>
AU25	38.5	80.1	77.4	<b>80.75</b>
AU26	20.1	52.4	52.6	<b>61.5</b>
AVG	29.64	53.32	56.38	59.43

\* AVG the average of 12 AUs F1-score

The bold denotes our method better than the others

method adopt randomly selector, in order to be consistent with our proposed method, the three methods would keep continuous once sampling every 4 frames. For a more rigorous comparison, we also set up an additional experiment using optical flow in combination with randomization.

On HDE protocol with test DISFA+ and CK+ datasets, our proposed method achieves 59.34 on F1, which is about 5.4% higher than the networks with the above AU-RCNN-LSTM methods in the 12 AUs. Even with the addition of randomization to Optical Flow RNN, it's still less accurate than our method. Obviously, the proposed OFS-GAN is more powerful than CNN-LSTM and AU-RCNN-LSTM, it can be better fused with the Optical Flow method to improve the performance of AUs classification.

## 5 Conclusion

In this paper, we realize that OFS-GAN can separate the subtle expression and head posture from face video, it is to detect the facial AUs changes. It can be encoded from the source frame to target frame of optical flow feature extraction and noise elimination by encoder. Combined with source frame and feature encoding, the synthetic image is generated by reverse decoding, and the expression detection of target frame can be realized. Our method has experimented by the HDE protocol achieved an F1-score of 0.5943 in the DAFSA+ and CK+ fuse dataset. According to the above our methods, improve the quality of AUs features and comparable analysis to most existing methods can be achieved. By experiments, it has shown itself to be a powerful tool for facial expression change detection in the video. Adopting optical flow feature extraction and adversarial, along with synthesis of fake pictures to adversarial variations with motion information, could be used to create an improved AUs analysis technique in the future. As our model is a general framework, it can be extended to other task like facial expression recognition or AUs expression translation, which are left as our future work.

## References

- Ekman, P., Friesen, W.V.: Manual for the Facial Action Coding System. Consulting Psychologists Press (1978)
- Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: a survey. *IEEE Trans. Affect. Comput.* **10**(3), 325–347 (2019). <https://doi.org/10.1109/TAFFC.2731763> (2017)
- Tong, Y., Chen, R., Yang, J., Wu, M.: Robust facial expression recognition based on local tri-directional coding pattern. In: Complex, Intelligent, and Software Intensive Systems—Proceedings of the 12th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS-2018, pp. 606–614 (2018)
- Zhi, R., Liu, M., Zhang, D.: A comprehensive survey on automatic facial action unit analysis. *Vis Comput.* **36**, 1067–1093 (2020). <https://doi.org/10.1007/s00371-019-01707-5>

5. Yang, H., Ciftci, U.A., Yin, L.: Facial expression recognition by de-expression residue learning. In: Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2168–2177 (2018)
6. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(6), 1113–1133 (2015). <https://doi.org/10.1109/TPAMI.2014.2366127>
7. Shan, L., Deng, W.: Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* **9** (2018)
8. Baltrušaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–6. IEEE (2015)
9. Romero, A., León, J., Arbeláez, P.: Multi-view dynamic facial action unit detection. *Image Vis. Comput.* (2017)
10. Chen, J., et al.: Learning person-specific models for facial expression and action unit recognition. *Pattern Recogn. Lett.* **34**(15), 1964–1970 (2013)
11. Liu, X., et al.: Adaptive metric learning with deep neural networks for video-based facial expression recognition. *J. Electron. Imaging* **27**(1), 406–414 (2018)
12. Hinz, T., Wermter, S.: Image generation and translation with disentangled representations. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2018)
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., WardeFarley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)
14. Xie, S., Hu, H., Chen, Y.: Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Trans. Circuits Syst. Video Technol.* **99**, 1 (2020)
15. Li, Y.: Self-supervised representation learning from videos for facial action unit detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2019)
16. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: a spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* **4**(2), 151–160 (2013)
17. Lucey, P., et al.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops. IEEE (2010)

# A Brief Survey on Privacy-Preserving Methods for Graph-Structured Data



Yunan Zhang , Tao Wu , Xingping Xian , and Yuqing Xu

**Abstract** As one of the main manifestations of big data, graph-structured data widely exists in various fields such as social networks, smart cities, medical health, and finance, and is characterized by high dimension, nonlinear, scale-free, small world, etc. Extensive graph-structured data provides sufficient data resources for scientific research and commercial applications. However, graph data mining not only reveals the intrinsic value of data, but also brings the risk of privacy disclosure. Therefore, how to protect graph-structured data privacy is of great significance. In this paper, we survey the very recent research development on privacy preserving methods for graph-structured data. We introduce the common sensitive information and the related privacy risks in graph data, and elaborate the background knowledge for privacy inference attack. Then, the privacy inference attack methods and privacy preservation methods for graph-structured data are summarized. Finally, the shortcomings of the current research about graph-structured data privacy preservation and the possible research directions are discussed.

**Keywords** Graph data · Privacy-preserving · Data anonymization

---

Y. Zhang  
Electric Power Research Institute, CSG, Guangzhou, China

Guangdong Provincial Key Laboratory of Power System Network Security, Guangzhou, China

T. Wu · X. Xian  
School of Cybersecurity and Information Law, Chongqing University of Posts and  
Telecommunications, Chongqing, China  
e-mail: [wutao@cqupt.edu.com](mailto:wutao@cqupt.edu.com)

Y. Xu  
School of Computer Science and Technology, Chongqing University of Posts and  
Telecommunications, Chongqing, China

## 1 Introduction

Over the past year, a great deal of research has been done to prove that the graph is a powerful tool for modeling the data with abundant relations. That is, graph can provide a very natural abstract representation of various data resources based on the collection of entities and the relations between them [1–3]. Specifically, in the field of computer vision, attributed graphs can be used to model and represent the local visual features in the image and the spatial relationship between them [4]. For natural language processing, the graph-structured representation can make full use of the semantic structure existing in all kinds of texts [5]. For signal processing, the sequential data can be represented based on the graph structure, where the nodes indicate the values at every time points and the relationships denotes the change of the values [6]. For knowledge service, knowledge graph can describe the connections between things in the world, so as to use graph algorithm to carry out knowledge reasoning and provide intelligent services [7]. Thus, graphs are ubiquitous in social networks, smart cities, e-commerce and other fields, and graph mining plays an essential role in social and economic activities.

Although the utilization of massive graph-structured data improves the development of society and economy, the mining process of graph-structured data leads to huge privacy risks [8, 9]. In fact, it is often possible to infer the user privacy by analyzing the released graph-structured data. To be specific, in the field of social network analysis, service providers may infer users' trust, living habits and political preferences based on graph-structured data [10, 11]. In the field of e-commerce, e-commerce platforms may infer users' income level, demographic attributes and other information [12]. In the field of urban transportation, map application providers can easily obtain individual living and working places [13].

Although researchers have studied the privacy-preserving problem in recent years and proposed various methods, such as k-anonymous, l-diversity, t-closeness and differential privacy, they are mainly developed for relational data [14]. The privacy protection methods for relational data only used the attribute values of each record as background knowledge, and they can not learn from the structure, relationship and node importance of graphs. Researchers have also proved that sensitive information can be inferred based on the relations between nodes in the graphs, thus graph-structured data is faced with great risk of privacy attack and disclosure [15, 16]. Therefore, the new privacy protection methods should be studied based on the intrinsic characteristics of graphs [17].

To sum up, based on the graph-structured data released by the data publishers, the malicious attackers may infer the hidden sensitive information of users. Once the sensitive information of users is stolen, it will bring serious adverse consequences to people's work, study and life. Therefore, it is urgent to study the privacy protection of graph-structured data. This paper focuses on the graph data privacy protection issues, systematically summarizes the graph data privacy protection methods proposed in recent years, analyzes the difficulties and challenges faced by graph data privacy protection, and prospects the future research direction.

## 2 Overview of Graph Data Privacy Preservation

### 2.1 Graph-Structured Data

In general, graph-structured data can be represented by a simple undirected graph  $G(V, E)$ , where node set  $V$  indicates the entities in the complex systems and edge set  $E$  denotes the relationships between the entities.

### 2.2 Graph Data Privacy

For the graph-structured data, its related privacy information can be discussed from the perspective of node-related privacy, attribute-related privacy and link-related privacy, as follows.

- (1) Node-related privacy. According to the structure and content characteristics of graph-structured data, node-related privacy can be divided into node existence, node identity and node structure.
- (2) Attribute-related privacy. Most of the graphs reflecting the complex systems have attribute information. If the attacker identifies the sensitive attribute that the user wants to hide, the sensitive attribute is considered to be leaked. Specifically, the attribute information can be subdivided into node attributes and link attributes.
- (3) Link-related privacy. In some cases, the links between nodes may be considered sensitive, such as transactional relationships, cooperative relationships, and the existence and the weight of them are not intended to be made public.

### 2.3 Graph Data Privacy Disclosure

In the context of big data, privacy refers to the sensitive information that can confirm the identity or characteristics of a particular individual but that individual does not want to be exposed. A privacy breach occurs when sensitive information is disclosed to an unauthorized attacker. At present, graph data is mainly faced with three kinds of privacy leakage risks, and the formal definitions are given as follows.

**Definition 1** (*Identity Privacy Disclosure*). Given graph  $G(V, E)$ , for each node  $v$  in the target node set  $V_t$ , the goal of node identity privacy disclosure is to infer its true identity based on its related links, neighbor structure, content attributes, etc.

**Definition 2** (*Attribute Privacy Disclosure*). Given graph  $G(V, E)$ , for node  $v$  in target node set  $V_t$ ,  $v \in V_t$ , attribute privacy disclosure aims to infer the attribute of node  $v$  by using its related links, neighbor structure and other information.

**Definition 3** (*Link Privacy Disclosure*). Given graph  $G(V, E)$ ,  $E'$  is defined as the set of sensitive links that are not willing to be exposed when data is released,  $E' \notin E$ . Suppose there is a random variable  $\hat{e}_{i,j}$  associated with the link between node  $v_i$  and  $v_j$ , the link privacy inference assigns a probability  $Pr(\hat{e}_{i,j} = \text{true})$  to the variable. If  $Pr(\hat{e}_{i,j} = \text{true} | \hat{e}_{i,j} \in E') > p$ , it is said that the private link will leak with the probability  $p$ .

## 3 Graph Data Privacy Inference Attack

### 3.1 Background Knowledge for Privacy Inference Attack

The background knowledge for privacy inference attack refers to the auxiliary information that the attacker can obtain and use for privacy attack in addition to the published data. The background knowledge of the attacker can be obtained through web crawlers, browser history traces, and exploring the same members of different websites. The background knowledge of the attacker can be divided into personal attributes, auxiliary network, structural information, etc. At the same time, the attacker can also combine a variety of background knowledge to attack privacy.

- (1) Personal attributes: Personal attributes describe the information related to users, such as name, address, age, and so on. Attackers can combine such information with other auxiliary information to carry out privacy inference attack.
- (2) Auxiliary graphs: Auxiliary graphs are those related to published graph data obtained by an attacker from other sources. For example, get a graph data with the same users and different relationships as the released graph from the outside world [18, 19]. Currently, the problem of node de-anonymization attacks using auxiliary networks has been widely studied, and related studies show that even auxiliary graphs with large noise can be used for sensitive information inference attacks [18].
- (3) Structural information: The attacker uses the topological information of the graph such as node degree, neighbor structure and subgraph as features to infer sensitive information. For example, in [20], node degree is used by attackers. In [21, 22], neighbor structure of nodes is used as background knowledge. In addition, the works [14, 23] used subgraph structure as background knowledge to carry out sensitive information re-identification attack.

### 3.2 Privacy Inference Attack Methods

In the graph privacy attack, the attacker can re-identify the sensitive nodes, links and attributes in the graph by analyzing and mining the publicly published anonymous graphs and background knowledge. The pioneering study on graph de-anonymization

was proposed by Backstrom et al. [15], who inferred whether there was a link between two specific nodes by using active and passive aggression methods. However, these two kinds of methods are only applicable to the graph data with naive anonymization for privacy protection. To this end, the follow-up researchers conducted a series of studies. Specifically, in terms of the background knowledge acquired by the attacker, the privacy inference attack can be categorized into node degree based attack, neighbor structure based attack, subgraph based attack and auxiliary graph based attack. According to whether the attacker owns the user's seed information, the attacks are divided into seed-based deanonymization and seed-free de-anonymization. Based on the objects being attacked, the attacks can be summarized as node identification attack, link inference attack and attribute inference attack. In this subsection, the existing privacy inference attack methods will be summarized from the view of attack target.

### (1) Node Identification Attack

In recent years, node identity attack becomes the main research point of graph data de-anonymization, and the existing works can be divided into two categories: attribute-based node identity attacks and structure-based node identity attacks. In the research of attribute-based node identity attack, most methods extract features from public files, such as user ID, location, etc., and extract features from published content, such as timestamp, geotag, etc., and then infer node identity [24, 25]. In the research of structure-based node identity attack, it is assumed that users have similar local structures in different networks, and the subgraph associated with the target node is used as the background knowledge of users. For example, Nilizadeh et al. [26] matched the anonymous graph with the auxiliary graph and identified the node identity through the community structure. Lee et al. [27] took the multi-hop neighbor information as the feature to match anonymous graph and auxiliary graph. Narayanan and Shmatikov [28] assumed the existence of seed nodes, and then propagated the mapping between seed nodes to other nodes to achieve de-anonymization. Ji et al. [29] conducted a comprehensive quantitative analysis of de-anonymization methods and proved the effectiveness of structural-based de-anonymization attacks.

### (2) Attribute Inference Attack

Attribute inference attack refers to the process in which an attacker infers missing or hidden sensitive attributes using acquired background knowledge [30]. The existing methods of attribute reasoning attack can be divided into relational attribute reasoning attack and behavior-based attribute reasoning attack. The relational-based attribute inference attacks make use of the theory of homogeneity [31], which assumes that friends are more likely to have similar attributes than strangers. The behavior-based attribute inference attack is the process of inferring a user's attributes from public information about the user's behavior and other public information similar to the user's behavior.

### (3) Link Inference Attack

Link inference attack, also known as link leakage or link re-identification, aims to identify hidden relationships between users in anonymized graphs. Ying et al. [14] studied sensitive relation protection and verified that similarity indexes play an important role in link privacy inference attacks. In order to verify the vulnerability of anonymity mechanism in graph data, vuokko et al. [32] proposed an inference method for randomized anonymous graph, assuming that the randomization method did not completely destroy the graph structure patterns, so as to reconstruct the original graph using eigen-decomposition. Fire et al. [33] believed that link prediction could be used for inference of hidden links and proposed an inference method based on machine learning classifier. Wu et al. [34] used the low-rank approximation method to reconstruct the original graph and proved that the disturbed link could be re-identified from the anonymous graph.

## 4 Graph Data Privacy Preservation

In the face of massive graph data resources, in order to achieve application development and scientific research, data resources need to be shared and analyzed. Data owners need to consider relevant privacy preservation strategies when sharing data to prevent leakage of sensitive information. To protect traditional rational data, many researchers proposed privacy preservation technologies such as k-anonymity, l-diversity, t-closeness, differential privacy. However, relational data is only a special case when the nodes in the graph are independent of each other. Due to the lack of consideration of the structure, relationship and importance of nodes in the graph, relational data privacy preservation cannot be directly applied to graph data. Therefore, traditional privacy preservation methods cannot meet the requirements of privacy preservation for graph data, and then researchers proposed naive anonymity, k-anonymity, random perturbation, clustering anonymity and other methods for graph data privacy preservation.

In this paper, graph data privacy preservation methods are mainly summarized into two categories: non-structural perturbation based privacy preservation and structural perturbation based privacy preservation. The former mainly include naïve anonymization [35, 36], which does not modify the graph structure, but replaces and eliminates some sensitive attributes in the graph data to achieve the purpose of privacy preservation. The latter privacy preservation methods include k-anonymity [37], graph perturbation [38], clustering [39], differential privacy [40] and so on. These privacy preservation mechanisms disturb the structure of the original graph to preserve the sensitive information.

## **4.1 Non-structural Perturbation Mechanism**

For the original graph, the data owner will anonymize the data that needs to be released. The most widely used method is to remove or change the sensitive information in graph data (such as identity numbers, social welfare numbers, and names), which is called as naïve anonymization. However, although the node identity information is deleted, the structure and other attribute information of original graph are still retained, and the attacker can still easily find the background knowledge of the target to identify it again. For example, the location of New York taxi drivers was leaked due to the simple naïve anonymization method [41]. Moreover, Many researchers protect sensitive information by perturbing the data value. That is, for data value perturbation, the data owner modifies the value to prevent attackers from obtaining certain key information. For instance, in some cases, a weighted graph can represent the frequency, cost, and similarity of interactions between two individuals [42]. Thus, data owners often reluctant to disclose such sensitive information, so they will perturb the edge weight value in order to protect such sensitive information before data publication. Liu [43] proposed two perturbation methods to protect the weight and path in the graph. One was Gaussian random increment method, which could protect the shortest path in a certain range. And another was Greedy algorithm that made the shortest path after perturbation the same as the original one.

## **4.2 Structural Perturbation Mechanism**

### **(1) k-anonymity**

In order to protect the graph nodes from being identified, the k-anonymity method is proposed, which is widely used in graph data privacy preservation. This kind of methods anonymize each node so that it is indistinguishable from at least  $k-1$  other nodes, so that the probability of privacy leakage is less than  $1/k$  [44]. Zou et al. [45] proposed a k-automorphism model to resist structural attack, and proposed an ID generalization technology for dynamic graph data to reduce the risk of ID disclosure in multiple data releases. Cheng et al. [46] proposed the k-isomorphism algorithm to resist link inference attack and extended it to dynamic networks. Liu et al. [47] studied degree-based attacks in graph data and proposed the k-degree anonymization (k-DA) algorithm.

### **(2) Structural Perturbation**

In structural perturbation, people often consider the relationship between users is sensitive information, such as friendships, hobbies, political preferences, and sexual orientation. Structural perturbation techniques can be used to protect these information by adding/deleting links or swapping links in the graph. Ying et al. [14]

proposed a link perturbation algorithm, which significantly improved the performance of link-based graph perturbation algorithm. And Ying et al. [14] also proposed Spctr Add/Del and Spctr Switch algorithms to protect spectrum characteristics in the graph. However, this method cannot quantify the degree of privacy protection, and there still exists privacy disclosure.

### (3) Cluster-based Anonymization

Cluster-based anonymization divides the nodes or links into several clusters based on attributes or structural characteristics, and then generalizes the nodes or edges in each cluster to achieve anonymization. Based on this idea, Hay et al. [48] proposed a clustering-based graph anonymized algorithm, which divided the original graph into clusters, and then generated super nodes and super edges according to the division. This method not only increases the uncertainty of the graph structure, but also reduces the availability of the graph. Recently, researchers have also proposed an anonymized method based on incremental clustering [49].

### (4) Differential Privacy Preservation

Differential privacy was firstly proposed in relational database, and in recent years, many works have extended it to graph data privacy preservation. To protect the sensitive links in graph, Sala et al. [40] firstly models the graph with  $dk$ -series, and then perturbed the  $dk$ -series to satisfy the  $\epsilon$ -differential privacy. Proserpio et al. [50] proposed a new graph publishing process to achieve differential privacy, which can reduce the noise scale by reducing the contribution of challenging records. Xiao et al. [51] used Hierarchical Random Graphs (HRG) to convert edges into link probabilities, and used the Markov chain Monte Carlo method to sample the HRG. Although differential privacy can protect data privacy well, it cannot be applied to large-scale graphs due to its complexity.

## 5 Discussion and Conclusions

In this paper, the problem of privacy preservation for graph-structured data is presented, the model of graph privacy preservation and attack is introduced, and the representative methods of graph privacy inference attack and privacy preservation proposed in recent years are summarized. Based on this paper, it can be found that most of the graph privacy preservation methods are mainly based on heuristics. The problem of graph privacy preservation is lack of theoretical modeling. At the same time, the consideration of graph data's characteristics are not sufficient in the privacy preservation methods. In addition, a large number of current research work focuses on the privacy protection of single-source data, while the attack and privacy preservation of graph data in multi-source scenarios have not been fully studied. Therefore, it is very necessary to conduct a more in-depth research on privacy preservation for graph-structured data.

**Acknowledgements** This work is partially supported by National Key R&D Program of China under Grant No. 2018YFB0904900, 2018YFB0904905; Natural Science Foundation of Chongqing under Grant No. cstc2020jcyj-msxmX0804, National Natural Science Foundation of China under Grant No. 61802039, 61772098, 61772091.

## References

1. Wu, T., Chen, Xian, X., Guo, Y.: Evolution prediction of multi-scale information diffusion dynamics. *Knowl. Based Syst.* **113**, 186–198 (2016)
2. Wu, T., Chen, L., Zhong, L., Xian, X.: Predicting the evolution of complex networks via similarity dynamics. *Phys. A* **465**, 662–672 (2017)
3. Wu, T., Guo, Y., Chen, L., Liu, Y.: Integrated structure investigation in complex networks by label propagation. *Phys. A* **448**, 68–80 (2016)
4. Zhang, Q., Song, X., Shao, X., et al.: Object discovery: soft attributed graph mining. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(3), 532–545 (2015)
5. Liu, B.: Natural language processing and text mining with graph-structured representations. <https://sites.ualberta.ca/~bang3/files/PhD-Thesis.pdf>
6. Cheng Z., Yang Y., Wang W., et al.: Time2Graph: revisiting time series modeling with dynamic shapelets. AAAI, pp. 1–9. AAAI Press, New York (2020)
7. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: representation, acquisition and applications. arXiv preprint arXiv: 2002.00388 (2020)
8. Beigi, G., Liu, H.: Privacy in social media: identification, mitigation and applications. Available: <https://arxiv.org/pdf/1808.02191.pdf> (2018)
9. Ji, S., Li, W., Srivatsa, M., et al.: General graph data de-anonymization: from mobility traces to social networks. *ACM Trans. Inf. Syst. Secur.* **18**(4), 1–29 (2016)
10. Beigi, G., Tang, J., Wang, S., et al.: Exploiting emotional information for trust/distrust prediction. In: SIAM International Conference on Data Mining, pp. 81–89. Florida (2016)
11. Gong, N., Liu, B.: Attribute inference attacks in online social networks. *ACM Trans. Priv. Secur.* **21**(1), 1–30 (2018)
12. Wang, P., Guo, J., Lan, Y., et al.: Your cart tells you: Inferring demographic attributes from purchase data. In: The Ninth ACM International Conference on Web Search and Data Mining, pp. 173–182. San Francisco (2016)
13. Mahmud, J., Nichols, J., Drews, C.: Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.* **5**(3), 1–21 (2014)
14. Ying, X., Wu, X.: Randomizing social networks: a spectrum preserving approach. In: SIAM International Conference on Data Mining, pp. 739–750. Society for Industrial and Applied Mathematics, Atlanta (2008)
15. Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In: The 16th International Conference on World Wide Web, pp. 181–190. New York (2007)
16. Korolova, A., Motwani, R., Nabar, S., et al.: Link privacy in social networks. In: The 17th ACM Conference on Information and Knowledge Management, pp. 289–298. New York (2008)
17. Xian, X., Wu, T., Wang, W., Wang, C., Xiao, Y., Liu, Y., Xu, G.: Towards link inference attack against network structure perturbation. *Knowl. Based Syst.* **218**, 106674 (2020)
18. Xian, X., Wu, T., Qiao, S., Wang, W., Liu, Y., Han, N.: Multi-view low-rank coding based structural de-anonymization for privacy preserving. *IEEE Access* **8**, 94575–94593 (2020)
19. Wondracek, G., Holz, T., Kirda, E., et al.: A practical attack to de-anonymize social network users. In: IEEE Symposium on Security and Privacy, pp. 223–238. California (2010)
20. Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: ACM Sigmod International Conference on Management of Data, pp. 93–106. New York (2008)

21. Zhou, B., Pei, J.: Preserving privacy in social networks against neighborhood attacks. In: IEEE 24th International Conference on Data Engineering, pp. 506–515. Cancun (2008)
22. Chester, S., Kapron, B., Ramesh, G., et al.: Why Waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. *Soc. Netw. Anal. Min.* **3**(3), 381–399 (2013)
23. Zou, L., Chen, L., Özsu, M.: K-automorphism: a general framework for privacy preserving network publication. *Proc. VLDB Endow.* **2**(1), 946–957 (2009)
24. Korayem, M., Crandall, D.: De-anonymizing users across heterogeneous social computing platforms. In: The 7th International IAAA Conference on Weblogs and Social Media, pp. 1–4. Boston (2013)
25. Shu, K., Wang, S., Tang, J., et al.: User identity linkage across online social networks: a review. *ACM SIGKDD Explor. Newsl.* **18**(2), 5–17 (2017)
26. Nilizadeh, S., Kapadia, A., Ahn, Y.: Community-enhanced de-anonymization of online social networks. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 537–548. Arizona (2014)
27. Lee, W., Liu, C., Ji, S., et al.: Blind de-anonymization attacks using social networks. In: The 2017 on Workshop on Privacy in the Electronic Society, pp. 1–4. Texas (2017)
28. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: The 30th IEEE Symposium on Security and Privacy, pp. 173–187. Washington (2009)
29. Ji, S., Li, W., Gong, N., et al.: Seed-based de-anonymizability quantification of social networks. *IEEE Trans. Inf. Forensics Secur.* **11**(7), 1398–1411 (2016)
30. Gong, N., Liu, B.: You are who you know and how you behave: attribute inference attacks via users' social friends and behaviors. In: 25th Usenix Security Symposium, pp. 979–995. Austin (2016)
31. McPherson, M., Smith-Lovin, L., Cook, J.: Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **27**(1), 415–444 (2001)
32. Vuokko, N., Terzi, E.: Reconstructing randomized social networks. In: SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 49–59. Ohil (2010)
33. Fire, M., Katz, G., Rokach, L., et al.: Links reconstruction attack. In: Security and Privacy in Social Networks. Springer, New York (2013)
34. Wu, L., Ying, X., Wu, X.: Reconstruction from randomized graph via low rank approximation. In: SIAM International Conference on Data Mining, pp. 60–71. Society for Industrial and Applied Mathematics, Ohil (2010)
35. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: Proceedings of the 30th IEEE Symposium on Security and Privacy, pp. 173–187. Oakland (2009)
36. Ji, S., Li, W., Srivatsa, M., Beyah, R.: Structural data de-anonymization: quantification, practice, and implications. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 1040–1053. New York (2014)
37. Casas-Roma, J., Herrera-Joancomart, J., Torra, V.: An algorithm for k-degree anonymity on large networks. In: Proceedings of the International Conference on Advances in Social Network Analysis and Mining, pp. 671–675. ACM Press, Niagara Falls (2013)
38. Ying, X., Wu, X.: Randomizing social networks: a spectrum preserving approach. In: Proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 739–750. Atlanta (2008)
39. Erlingsson, Ú., Pihur, V., Korolova, A.: Rappor: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 1054–1067. Scottsdale (2014)
40. Sala, A., et al.: Sharing graphs using differentially private graph models. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 81–98. Berlin (2011)
41. Ji, S., Li, W., Srivatsa, M., et al.: Structural data de-anonymization: quantification, practice, and implications. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 1040–1053. Arizona (2014)

42. Ji, S., et al.: Secgraph: a uniform and open-source evaluation system for graph data anonymization and de-anonymization. In: 24th USENIX Security Symposium, pp. 303–318. Washington (2015)
43. Liu, L., et al.: Privacy preserving in social networks against sensitive edge disclosure. Technical report technical report CMIDA-HIPSCCS 006-08. Department of Computer Science, University of Kentucky, KY (2008)
44. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncert. Fuzziness Knowl. Based Syst.* **10**(5), 557–570 (2002)
45. Zou, L., Lei, C., Özsu, M.T.: K-automorphism: a general framework for privacy preserving network publication. *Proc. VLDB Endow.* **2**(1), 946–957 (2009)
46. Cheng, J., Fu, A., Liu, J.: K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 459–470. Indianapolis (2010)
47. Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of data, pp. 93–106. Vancouver (2008)
48. Hay, M., Miklau, G., Jensen, D., et al.: Resisting structural re-identification in anonymized social networks. *Proc. VLDB Endow.* **1**(1), 102–114 (2008)
49. Tassa, T., Cohen, D.J.: Anonymization of centralized and distributed social networks by sequential clustering. *IEEE Trans. Knowl. Data Eng.* **25**(2), 311–324 (2011)
50. Proserpio, D., Goldberg, S., McSherry, F.: Calibrating data to sensitivity in private data analysis: a platform for differentially-private analysis of weighted datasets. *Proc. VLDB Endow.* **7**(8), 637–648 (2014)
51. Xiao, Q., Chen, R., Tan, K.: Differentially private network data release via structural inference. In: The 20th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 911–920. New York, United States (2014)

# Disentangled Representation Learning from Videos for Facial Action Unit Detection



Zhe Dong , Huijuan Zhao , Jing Juan , Shuangjiang He , and Zhi Tao

**Abstract** Facial Action Unit (AU) detection is a challenging task for the reason that AU features extracted from videos always entangle other inevitable variations, including head posture motion characteristics, which is even much more intense than facial actions, and individual facial features because of race, age, gender or face shape. These AU-unrelated features would adverse the AU detection performance. To achieve better performance of AU detection, we proposed a novel Feature Disentangled Autoencoder (FDAE) to learn more discriminative facial action representation from large amounts of videos. Different from previous approaches, FDAE disentangled AU-related features, head posture motion characteristics and identity code characteristics to eliminate the impact of irrelevant factors. At the same time, we added two classifiers to discriminate the AU embedding and identity code embedding respectively to accelerate training process and make the model more stable and robust. Experiments on BP4D and DISFA demonstrated that the learned representation is discriminative, where FDAE outperformed or was comparable with existing representation learning method for AU detection.

**Keywords** Action units (AUs) · Facial action unit detection · Representation learning · Feature disentangled autoencoder (FDAE)

---

Z. Dong · Z. Tao

Wuhan Fiberhome Integration Technologies Co., Ltd., Wuhan, China

H. Zhao · S. He

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China

e-mail: [D202080766@hust.edu.cn](mailto:D202080766@hust.edu.cn)

J. Juan

Wuhan Intelligent Medica Research Institute Co., Ltd., Wuhan, China

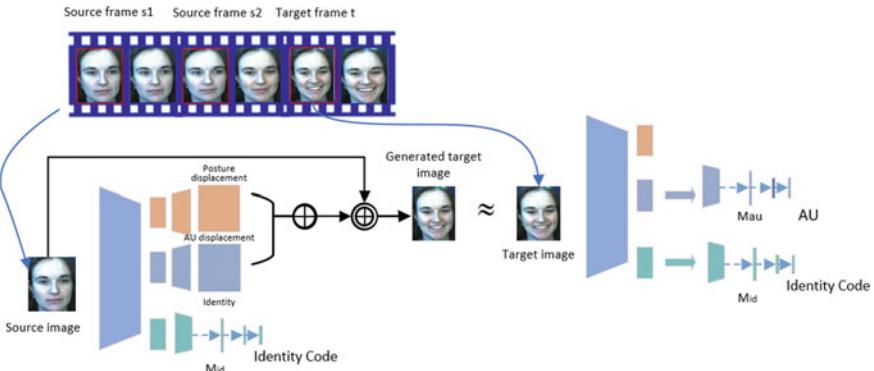
## 1 Introduction

Facial expressions indicate momentary changes in facial appearance during emotional communication and reveal one's thoughts, feelings and internal psychological states. Facial Action Units (AUs) are known as the basic building blocks in formulating various facial expressions. Therefore, successful AU detection would make a great significance for complicated expression recognition. Automatic AU detection has been investigated for decades as one of the most vital research topics in facial expression analysis. Traditional approaches of AU detection focus on the correlations between AU detection and facial landmarks [1] or integrating extra information like optical flow [2] to represent the movements in face to get more precise features for subtle motions of facial local region. With the development of deep learning, a great number of deep learning methods [3–5] are proposed to learn deeper and more discriminative representation for AU detection. In fact, those methods have high error rate because most approaches use full face image as input and might learn unreliable representation, which would lead to not using the correct context information for the detection task. For instance, since different people have great individual differences because of race, age, gender or face shape, the detection system might not use the AU-related features rather incorrect inherent appearance context information of the individual face. We define the individual appearance context information as identity code. On the other hand, [6] divided facial images into uniform patches to learning the most ones for AU detection, [7] proposed upper, middle and lower regions to detect AU respectively and integrate the results in the end to capture facial action features in local region. However, the region selection of these methods is not based on AU region knowledge, which might cause the model to fail to learn precise AU association information. What's more, the representations learned from aforementioned methods suffer from head posture motions characteristics, which might be even more intense than facial actions and would mislead the models to use AU-unrelated features to predict AU or even adverse the performance of the detection system.

In this paper, we proposed a Feature Disentangled Autoencoder (FDAE) to learn discriminative facial action representation for AU detection based on disentangling AU-related features, head posture motions characteristics and identity code characteristics from large amounts of videos. At the same time, we added AU classifier to supervise AU embedding learning and identity code classifier to supervise identity code embedding learning. In addition, the two classifiers would stable and speed up the training process, and make the model more robust. Figure 1 illustrates the main idea of FDAE.

In summary, compared with existing approaches, the main contributions of this paper are as follows:

- (1) We proposed a novel Feature Disentangled Autoencoder (FDAE) framework to learn facial representation for AU detection based on disentangling AU-related features and head posture motions characteristics and identity code characteristics from large amounts of videos.



**Fig. 1** The main idea of Feature Disentangled Autoencoder (FDAE) for AU detection

- (2) We proposed a feature disentangling module to encode images and disentangle AU-related feature from other unrelated facial action features, which makes the representations learned from FDAE more robust.
- (3) Experiments demonstrated that learned representation is discriminative for AU detection, where FDAE outperforms or is comparable with existing representation learning method for AU detection.

## 2 Relate Work

Facial expression is one of the most powerful, natural and universal signals to reveal one's thoughts, feeling and internal psychological states. The expressions appear in local facial region by the way of facial muscle motions [8]. Facial Action Coding System (FACS) [9] decomposes facial expressions into 44 unique sets of atomic nonoverlapping facial muscle actions called action units (AUs) and plays a fundamental role in accurately characterizing facial actions.

Automatic AU detection has been a growing field that has attracted attention of many researchers [10]. To achieve more discriminative features for AU detection, researchers have proposed a good deal of approaches in traditional manner [11] proposed method based on Histograms of Oriented Gradients (HOG) to extract feature of local or global changes of facial components [12] portrays the distance or direction of salient facial skins based on Optical Flow. With the development of deep learning, convolutional neural network has been widely adopted for AU detection [11] to make the feature more discriminative. Zhao et al. [5] proposed the Joint Patch and Multi-label Learning (JPML) method for AU detection. Happy et al. [6] adopted grid region to select salient region for AU detection. Xia et al. [7] proposed approach based on dividing the face into upper, middle and bottom regions, and passed feature maps into corresponding model to predict AU labels respectively, and fuse three results to receive final result. However, most of these approaches used

**Table 1** Rules for defining AU centers

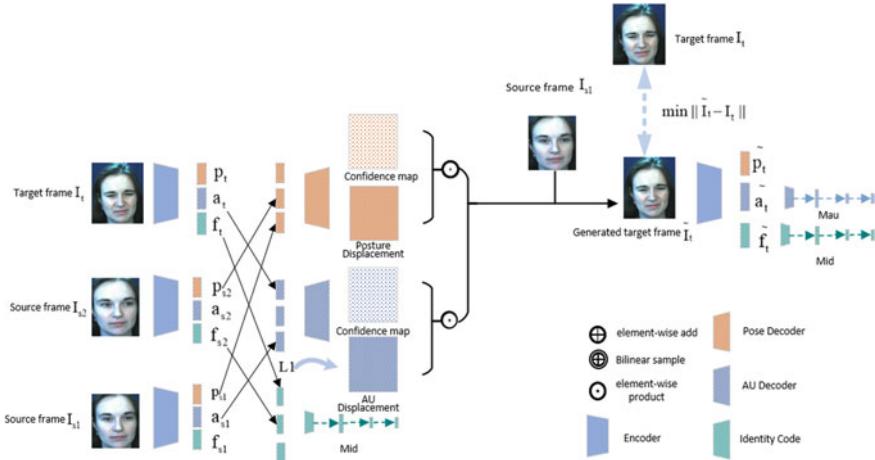
AU index	Au name	Muscle name
1	Inner brow raiser	Frontalis
2	Outer brow raiser	Frontalis
4	Brow lowerer	Corrugator supercilii
6	Cheek raiser	Orbicularis oculi
7	Lid tightener	Orbicularis oculi
10	Upper lip raiser	Levator labii superioris
12	Lip corner puller	Zygomaticus major
14	Dimpler	Buccinator
15	Lip corner	Triangularis
17	Depressor	Mentalis
23	Chin raiser	Orbicularis oris
24	Lip pressor	Orbicularis oris

full face image as input and consider the AU-unrelated feature as key for AU detection, which leads to not using the correct context information in detection system, or divided facial region into local region, leaving out of sufficient consideration of AU relations. More important, these methods utilized samples within head posture variations and might consider individual differences as key features for AU detection.

This paper proposed a Feature Disentangled Autoencoder (FDAE) to detect AUs by disentangling AU-related features, head posture characteristics, and identity-related characteristics. Experiments on BP4D and DISFA demonstrated that the learned representation is discriminative, where FDAE outperformed or was comparable with existing representation learning method for AU detection (Table 1).

### 3 Proposed Method

In this paper, we proposed a Feature Disentangled Autoencoder (FDAE) to learn discriminative facial action representation for AU detection based on disentangling AU-related features, head posture motions characteristics and identity code characteristics from large amounts of videos. Figure 2 illustrates the main idea of FDAE. FDAE consists of feature disentangling module, target reconstruction module, AU classifier and identity code classifier. In Feature disentangling module, FDAE learns facial representation by disentangling AU movements, head posture movements, and identity code characteristics from two source face images and one target face image respectively. Then, FDAE uses identity code classifier to supervise identity code learning. In target reconstruction module, FDAE concatenates AU movements and head posture movements of three face images respectively to generate the grid sampler to generate an approximative target facial image from source facial image



**Fig. 2** The main idea of FDAE. Given two source images  $I_{s1}, I_{s2}$  and a target image  $I_t$ , the encoder E passes them separately into pose-related feature embedding  $[p_{s1}, p_{s2}, p_t]$ , AU-related feature embedding  $[a_{s1}, a_{s2}, a_t]$  and facial identity code embedding  $[f_{s1}, f_{s2}, f_t]$ . The embeddings are decoded into pose-related displacement and au-related displacement by their corresponding decoders respectively

called as generated target image. Classifiers are used to predict the AU labels and identity code of the generated target image respectively and make FDAE more stable and robust.

### 3.1 Feature Disentangling Module

As is shown in Fig. 2, we randomly selected  $I_{s1}$  from the trained expression video nearby the onset frame as approximate onset frame because it is difficult to find accurate onset frame in practical emotional video when people are talking with different expressions. For the similar reason,  $I_t$  was randomly sampled from the video nearby the apex frame as approximate apex frame. The middle frame named  $I_{s2}$  was sampled from the video of 25–75% to enhance the representation of FDAE. FDAE respectively passed the three frames into a same encoder E, and generated their corresponding AU-related feature embeddings  $[a_{s1}, a_{s2}, a_t]$ , head posture features embeddings  $[p_{s1}, p_{s2}, p_t]$ , and facial identity code embeddings  $[f_{s1}, f_{s2}, f_t]$ . After that, FDAE concatenated the head posture feature embeddings as head posture embedding and passed it into head posture decoder  $D^P$  to generate head posture displacements, which denoted how and where the head posture of the source image was changed into the target image. In the meantime, FDAE generated head posture confidence map, depicted as the offset between the source image and the target image named posture-related displacement, to denote how believable the change of the pixel and

the movements caused by head posture were. Similarly, FDAE concatenated AU-related feature embeddings as AU embedding and passed it into AU-related decoder  $D^a$  to generate AU confidence map and AU-related displacement.

For the input image with shape (W, H), the pose-related displacements were formulated as a matrix of vectors  $T^p$ .  $T_{xy}^p = \{\delta_x, \delta_y\}$  denoted the offset of source image  $I_{s1}$  at position (x, y). Similarly, the AU-related displacements were formulated as  $T^a$ . As a matter of experience, AU-related displacements were much weaker than head posture displacements and only led to facial local region movements of one or a group of muscles, we added L1 regularization on  $T^a$  to keep the AU-related displacements subtle and sparse, which was formulated as:

$$L_1^a = \sum \|T_{xy}^a\|_1, \quad (1)$$

where x, y denoted all the locations in the face images. We minimize L1 regularization and enforce the AUs movement to be subtle.

### 3.2 Target Reconstruction

FDAE learned the AU-related features, head posture characteristics and identity code characteristics through predicting AU-related displacements, head posture displacement and identity code. In order to ensure the representational capacity of the displacement features, we proposed target reconstruction module to generate an approximative target facial image from source facial image by integrating the AU-related displacement  $T^a$  and pose-related displacement  $T^p$  and their corresponding confidence map  $\partial^a \in \mathbb{R}^{H \times W}$  and  $\partial^p \in \mathbb{R}^{H \times W}$ . The confidence map denoted how confident of the displacement to represent the movement features, that is, if the source image's head posture was very different from the target image, the confidence map would have a high certainty. FDAE reconstructed the target image by linearly combining two displacement, which was formulated as Formula (2), and used the result as a grid sampler to sample the onset frame to produce the approximate target face image.

$$\tilde{I}_t = I_{s1} \cdot \left( \partial_{xy}^a T_{xy}^a + \partial_{xy}^p T_{xy}^p \right) = T(I_s), \quad (2)$$

where  $T_{xy}^a, T_{xy}^p \in R^2$  were the pixel displacement vectors of location (x, y) and  $\tilde{I}_t$  denoted the reconstructed image called as generated target image from the source image,  $\partial_{xy}^a, \partial_{xy}^p \in R$  were the confidence of location (x, y).

We considered T as the transformation mapping the source image  $I_s$  into the target image  $I_t$ , therefore, we should minimize the discrepancy between  $\tilde{I}_t$  and  $I_t$ . We calculate the reconstruction loss as:

$$L_{rec} = \|\tilde{I}_t - I_t\|_1, \quad (3)$$

where  $I_t$  denotes the sampled target image. On the other hand, we reconstructed the target image to approximate apex frame, the embeddings of  $\tilde{I}_t$  should be similar to the ones of  $I_t$ . That is to say, the target image and the generated target image should have similar AU-related embedding, head posture features embedding and identity code embedding. We formulated the similarity by minimizing the embedding loss:

$$L_{emb} = \|\tilde{p}_t - p_t\|^2 + \|\tilde{f}_t - f_t\|^2 + \|\tilde{a}_t - a_t\|^2, \quad (4)$$

where  $\tilde{p}_t$ ,  $\tilde{f}_t$  and  $\tilde{a}_t$  were the embeddings of the reconstructed image passed into the encoder E,  $p_t$ ,  $f_t$  and  $a_t$  were the embeddings of the target image passed into the same encoder E.

The generated images from FDAE were visually meaningless at the beginning of training but treated as positive cases and contributed the most of the loss. Since the loss would mislead the reconstruction module to a wrong learning direction, we added AU classifier and identity code classifier to stabilize the training of the model. In the end, we trained the model with progressively difficult samples according to the curriculum learning [13] strategy. The identity code classifier predicted the identity code of the subject, we adopted softmax cross-entropy loss function, as shown in Formula (4).

$$L^{id}(y, \tilde{y}) = -\frac{1}{N} \sum_{n=1}^N \{y \log(\tilde{y})\}, \quad (5)$$

where  $y$  was the identity code of the sample, and  $\tilde{y}$  is the prediction probability. The AU classifier predicted multiple AU labels, we adopted multi-label sigmoid cross-entropy loss function, formulated as Formula (5)

$$L^{au}(y, \tilde{y}) = -\frac{1}{N} \sum_{n=1}^N \sum_{l=1}^L \{y_{n,l} \log(\tilde{y}_{n,l})\}, \quad (6)$$

where  $y_{n,l}$  is the AU label with multiple labels of the sample, and  $\tilde{y}_{n,l}$  is the prediction probability.

## 4 Experimental Evaluation

We have performed a series of experiments on two major AU datasets to evaluate the effectiveness of Fdae on AU detection task.

## 4.1 Datasets

AU datasets are harder to obtain, because labeling AUs are complicated and time consuming. It would take one at least 100 h to be a FACS expert. Even for the experts, manually code an AU for 2 min video would take 1 h or more. BP4D [14] and DISFA [15] are two main dynamic AU datasets for AU detection tasks, which are also the ones we use in our experiments.

**BP4D:** There are 328 videos with 23 female and 18 male young adults. The samples are obtained when the subjects are watching videos with spontaneous emotional expressions. AUs are coded by FACS coders with labels and intensity of 12 AUs. The intensities are scales from 0 to 5, where 0 means the absence of an AU, and scale of 1–5 represents minimum to maximum intensity of each AU. We can obtain around 140,000 images with AU label.

**DISFA:** DISFA consists of 593 spontaneous facial actions sequences from 27 adult subjects. As similar to BP4D, every frame in DISFA is coded by FACS coder and the intensity of 12 AUs in 0–5. There are more than 100,000 AU-labeled images in DISFA.

The frames with intensities greater than 1 were treated as positive samples while others were considered as negative samples in our experiment.

## 4.2 Experimental Setting

AU detection is a multi-label task to detect whether AUs present in the face. First of all, as the distribution of all AU datasets are highly imbalanced, we inversely reweighted the samples of the under-represented categories based on the class frequencies. And Then we partitioned the dataset into 3 folds based on the subjects. Two of the folds were used as training set and the third one was used for testing. We resized the face images into  $256 \times 256$  and normalized them after cropping face area from frames of video and adopted Leave-One-Subject-Out (LOSO) cross-validation to ensure subject-independent evaluation. The detection is a challenging task because several AUs always don't present independently. At the same time, AU detection model might not learn the features impartially based on imbalanced datasets. Since F1 score can better describe the performance of an algorithm especially when samples are imbalanced, we adopted F1 score as then evaluation metric, which is defined as  $F1 = 2PR/(P + R)$ , where P and R denote precision and recall respectively. What's more, we calculated the average over all AUs to assess the overall performance of our FDAE.

The experiments were all trained on 3080Ti GPU using same hyper parameters: BATCH = 32, EPOCH = 100, and the optimizer is Adam with an initial learning rate of  $1e-4$ . The image pairs randomly sampled during training might contain large deviations in the beginning, so we adopted the curriculum learning [15] strategy to

use progressively difficult samples to train the model. Specifically, we only selected top 50% of the sorted losses of the samples in ascending order within a batch to back-propagate. When the validation saturated, we would change the ranking to between top 10 and 60%, and so on. Full loss is formulated as:

$$L = \frac{\sigma 1}{W * H} L_{rec} + \frac{\sigma 2}{W * H} L_1^a + \frac{\sigma 3}{W * H} L_{emb} + \sigma 4 L^{id} + \sigma 5 L^{au}, \quad (7)$$

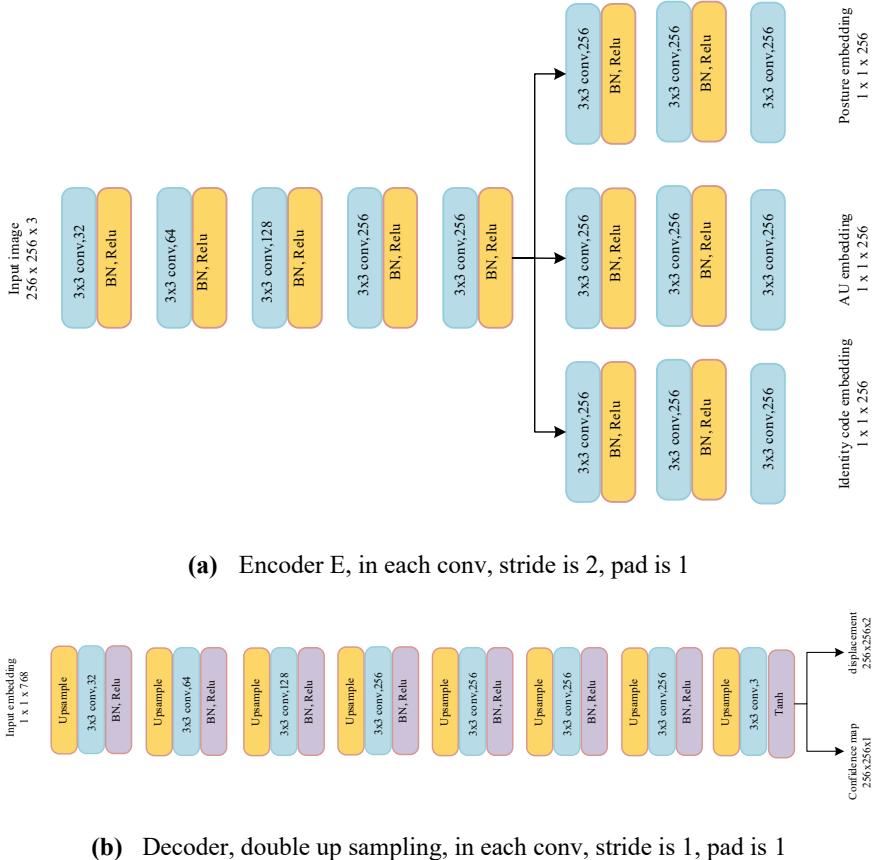
where W and H denoted the weight and height of input image. The weight of each loss  $\sigma 1, \sigma 2, \sigma 3, \sigma 4$  and  $\sigma 5$  are set as 0.01, 0.1, 0.1, 0.1, 0.1.

### 4.3 Structures of the Encoder and Decoder

Figure 3 demonstrated the structure of the encoder E and the decoder of FDAE in our experiments. The encoder E shown in Fig. 3a contained a shared shallow backbone network followed by three parallel branches to transform the input image in size of  $256 \times 256 \times 3$  into AU-related features embedding, posture motion embedding and identity code embedding in size of  $1 \times 1 \times 256$  respectively. We choose shared backbone because shallow level could represent general basic level features of most input images such as texture. Figure 3b shows the structure of decoder. The input of the decoder is concatenation of corresponding features, i.e., the concatenation of posture motion features from three face images. The decoder outputted displacement feature map in the size of  $256 \times 256 \times 2$  and confidence map in size of  $256 \times 256 \times 1$ .

### 4.4 Result and Analysis

To demonstrate the effectiveness of FDAE, we compared with some other methods, such as a baseline linear SVM (LSVM) [4], JPML [5], DRML [16] and so on, including handcrafted features and automatic features. The result in Tables 2 and 3 showed that FEAЕ outperformed or was comparable with existing representation learning method for AU detection, which suggested that the disentangled AU features were relatively robust and effective. However, not all AU F1 score are better than other methods, might because the AU distribution of the datasets are imbalanced and AU are not same intensity. At the same time, we trained our model only on a few datasets, which might limit the representation and generalization of FDAE in more extensive application.



**Fig. 3** The structure of encoder and decoder

## 5 Conclusion

This paper proposed a novel end-to-end Feature Disentangled Autoencoder (FDAE) to acquire a more discriminative facial action representation for AU detection tasks. Different from previous models, FDAE learned to disentangle facial action features from head posture motion characteristics and identity code characteristics to yield a better detection performance. Since FDAE could disentangle features of images and fuse the separated features with other small scene data sets, we could try to transfer facial action into other facial images, especially in the application scenario of data set shortage. Experiments conducted on BP4D and DISFA demonstrated that learned representation is discriminative for AU detection, where FDAE received average F1 score of 52.62 on BP4D dataset and 48.08 on DISFA. FDAE outperformed or was comparable with existing representation learning method for AU detection.

**Table 2** F1 score on BP4D dataset

AU	LSVM	JPML [4]	DRML [5]	FVGG	E-Net	FDAE (our)
1	23.2	32.6	36.4	27.8	<b>37.6</b>	37.3
2	22.8	25.6	<b>41.8</b>	27.6	32.1	34.87
4	23.1	37.4	43	18.3	<b>44.2</b>	38.5
6	27.2	42.3	55	69.7	75.6	<b>76.67</b>
7	47.1	50.5	67	69.1	<b>74.5</b>	73.24
10	77.2	72.2	66.3	78.1	80.8	<b>80.92</b>
12	63.7	74.1	65.8	63.2	<b>85.1</b>	78.81
14	64.3	<b>65.7</b>	54.1	36.4	56.8	59.32
15	18.4	<b>38.1</b>	33.2	26.1	31.6	33.9
17	33	40	48	50.7	55.6	<b>55.7</b>
23	19.4	30.4	<b>31.7</b>	22.8	21.9	29.3
24	20.7	<b>42.3</b>	30	35.9	29.1	32.85
Avg	35.3	45.9	48.3	43.8	52.1	<b>52.62</b>

**Table 3** F1 score on DISFA dataset

AU	LSVM	APL	DRML [5]	FVGG	ROI	FDAE(our)
1	10.8	11.4	17.3	32.5	<b>41.5</b>	39.23
2	10	12	17.7	24.3	<b>26.4</b>	26.21
4	21.8	30.1	37.4	61	<b>66.4</b>	<b>39.31</b>
6	15.7	12.4	29	34.2	50.7	<b>52.31</b>
9	11.5	10.1	10.7	1.67	8.5	<b>16.77</b>
12	70.4	65.9	37.7	72.1	<b>89.3</b>	80.13
25	12	21.4	38.5	87.3	<b>88.9</b>	87.31
26	22.1	<b>26.9</b>	20.1	7.1	15.6	24.4
Avg	21.8	23.8	26.7	40.2	<b>48.5</b>	48.08

As our model could generate images based on AU-related displacement and pose-related displacement, we would apply our FDAE to expand dataset for the tasks with inadequate datasets.

## References

1. Liu, P., et al.: Multi-modality empowered network for facial action unit detection. In: WACV (2019)
2. Romero, A., León, J., Arbeláez, P.: Multi-view dynamic facial action unit detection. ArXiv abs/1704.07863 (2017)
3. Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: a region-based deep enhancing and cropping

- approach for facial action unit detection. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 103–110. IEEE (2017)
- 4. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: a library for large linear classification. *JMLR* **9**, 1871–1874 (2008)
  - 5. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2207–2216 (2015)
  - 6. Happy, S.L., Routray, A.: Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Aff. Comput.* **6**(1), 1–12 (2015). <https://doi.org/10.1109/TAFFC.2014.2386334>
  - 7. Xia, Y.: Upper, middle and lower region learning for facial action unit detection. *ArXiv abs/2002.04023* (2020)
  - 8. Freitas-Magalhães, A.: Facial expression of emotion: from theory to application. Leya (2013)
  - 9. Ekman, P., Friesen, W.V.: Facial Action Coding System. Palo Alto, California: Consulting Psychologists Press (1978)
  - 10. Li, Y., Zeng, J.B., Liu, X., Shan, S.G.: Progress and challenges in facial action unit detection. *J. Image Graph.* **25**(11), 2293–2305 (2020)
  - 11. Valstar, M.F., Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. *IEEE Trans. Syst. Man Cybernet. Part B Cybernet.* **42**(1), 28–43 (2012)
  - 12. Romero, A., León, J., Arbeláez, P.: Multi-view dynamic facial action unit detection. In: Image and Vision Computing (2017)
  - 13. Bengio, Y., Louradour, J., Ronan Collobert, T., Weston, J.: Curriculum learning. In: ICML. ACM (2009)
  - 14. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4ds spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* **32**(10), 692–706 (2014)
  - 15. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: a spontaneous facial action intensity database. *IEEE Trans. Aff. Comput.* (2012)
  - 16. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multilabel learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3391–3399 (2016)

# Facial Expression Recognition Based on Images Captured and Refined with Synchronized Voice Activity Detection



Xiaoqing Jiang , Lingyin Wang , and Yue Zhao

**Abstract** Facial expression recognition from video is important in affective computing field. Because facial expression is a slowly changing process and often accompanied by other emotional psychological behaviors, there are a lot of redundant data in facial expression images obtained directly from the video. The redundancy results in the performance degradation in the learning of features and training of facial expression recognition model. In this paper, a capturing and refining method with synchronized Voice Activity Detection (VAD) is adopted to emotional videos in order to obtain facial expression images with better emotional recognizability. The refined images and the pre-trained ImageNet-VGG-F model are utilized to achieve high performance on the facial expression recognition. Videos in RAVDESS database are used in the experiments, and the recognition accuracy is 91.667% for the specific speaker, which is 19.5% higher than the experiments without synchronized VAD. The experimental results show that this method proposed in this paper is effective.

**Keywords** Facial expression recognition · Voice activity detection · Convolutional neural networks

## 1 Introduction

Emotion information is essential in communication in our daily life, and emotion recognition is a hot topic in the research of Artificial Intelligence (AI) and Human Computer Interaction (HCI). The effective processing of emotion also can bring great breaks to traditional technologies, which is also the main research motivation in affective computing field.

Emotion information can be carried by various signals including behavior signals and physiological signals, such as speech, facial expression image, move pattern, bait, heart rate, blood pressure, and EEG. Compared to physiological signals, the behavior emotional signals are easier to be sensed and captured. Take emotional

---

X. Jiang · L. Wang · Y. Zhao

School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong, China

e-mail: [ise\\_jiangxq@ujn.edu.cn](mailto:ise_jiangxq@ujn.edu.cn)

speech and facial expression images as example, they are the most common modals of emotion that can be captured from videos recorded by smart phones, computer cameras, webcams, or monitoring equipment. Thus, a lot of emotion recognition effort is based on speech signals or facial expression images.

Research shows that complementary information exists in multi-modal emotional signals, which is important to the improvement of performance and robustness of the emotion recognition system [1–3]. Emotional recognition has also achieved a lot of valuable research results recently based on multi-modal signals [4–6]. However, there are inevitable difficulties in the processing of multi-modal emotion signals and the training of corresponding model in multi-modal emotion recognition because of the high dimension of the multi-modal feature set and complexity of computation. On the other hand, data redundancy also exists during the multi-modal feature learning because of the strong relationship among multi-modal emotional data. So it's hard to say that multi-modal emotion recognition must be superior to single-modal research. To achieve the similar recognition performance as multi-modal signal by using single-modal data with better emotional recognizability is one of the challenges in emotion recognition field.

In this paper, emotional videos that can be separated into audio and visual modals are adopted in the research. There are silence segments and speech segments in the audio modal. Similarly, the corresponding facial expressions without voice are usually ‘static’, i.e., most of them stay the same, which decreases the emotional recognizability of the visual modal. This problem often is ignored in most of the audio-visual emotion recognition.

Refining the facial expression images of the visual modal can combine with the detection of speech activity. Voice Activity Detection (VAD) is a common preprocessing procedure in speech signal processing, and it can detect the silence segments and the speech segments. In this paper, according to the results of VAD, only the facial images that are synchronous with speech segments will be captured. This step can be viewed as an effective selection or refining to the facial expression images before learning and training, because only the images with better emotional recognizability will be captured and the redundant images in silence segments are deleted.

As a pattern recognition problem, a variety of methods can be adopted in facial expression recognition field, such as Support Vector Machine (SVM), Hidden Markov model (HMM), and Artificial Neural Network (ANN), etc. Deep learning is the most popular and effective method today [7]. Deep learning is a hot research topic and the architectures such as deep neural networks, deep belief networks, and convolutional neural networks have been applied to various fields including computer vision, speech recognition, natural language processing, medical image analysis [8, 9]. Convolutional Neural Network (CNN) is a typical architecture of deep neural networks, it has wide application in image and video recognition, image classification, medical image analysis, and natural language processing. Transfer learning is a machine learning technique where a model trained on one task is repurposed on a second related task. We can also adjust the trained convolutional layers to suit new problems, which is fine tuning [10, 11]. In deep learning field, the pretrained model together with transfer learning and fine tuning can improve the efficiency of learning

task effectively. In this paper, the pretrained ImageNet-VGG-F model is adopted in facial expression recognition.

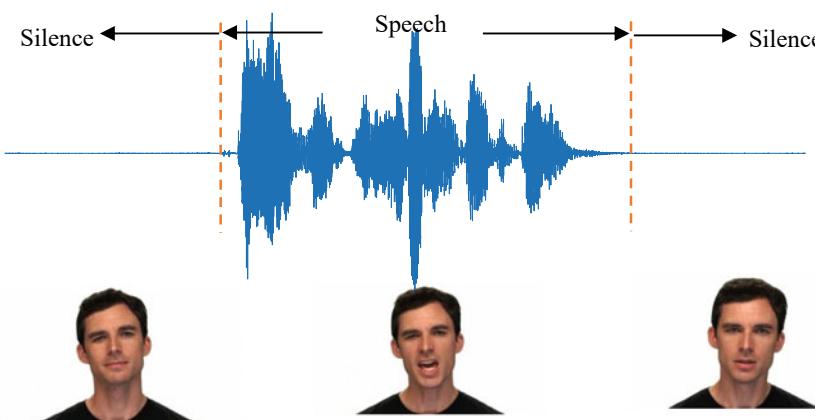
The content of this paper includes: in Sect. 2, redundancy in facial expression images is analyzed, and the double thresholds VAD algorithm is introduced and adopted in the capturing and refining of facial expression images; Sect. 3 is about the pretrained CNN model in facial expression recognition and the introduction of the research diagram; Sect. 4 is about experiments and results analysis; and finally, conclusions and the future work will be given in the Sect. 5.

## 2 Redundancy Analysis and the Refined Capturing of Facial Expression Images

The expression of human's face is relatively slowly comparing to the frequency of signals. Usually, the facial expression can last 0.5–4 s when the single emotions occur without modification or change [12]. As mentioned above, audio and visual modals also have strong relationship. Though sometimes we can capture emotional facial expression images in silence or have emotional voice with neutral facial expression, in most of the situation, our facial expression will change more obviously along when the speaker speeches.

### 2.1 Components in Speech

Except for voiced and unvoiced speech segments, there are silence segments in the audio file. These different components in an audio signal are shown in Fig. 1. Figure 1



**Fig. 1** Facial expression intensity of different speech segments

shows the images captured from an angry video of RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database and the synchronous silence and speech segments respectively.

The RAVDESS database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in North American accent. 8 emotions including calm, happy, sad, angry, fearful, surprise, disgust and neutral are contained, and each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression [13]. As in illustrated in Fig. 1, different segment has different intensity of facial expression. In the silence segments at the beginning and ending part, the facial expression is often in a static situation, where the emotion is hard to be recognized accurately even by our human beings. Facial images captured in silence segments aren't proper to be used to learn the features and train the recognition model.

## 2.2 Double Thresholds VAD and the Result

In order to solve the above problem, we adopt VAD algorithm based on the combination of short-time energy  $E_n$  and short-time zero crossing rate  $Z_n$  to the audio signal synchronous with the visual modal. The algorithm is named the double thresholds VAD because it uses thresholds of these two parameters respectively. Speech frames can be obtained from the audio modal with windows whose length is about 30 ms. The short time energy and the short time zero crossing rate are calculated for every frame.

If the  $n$ th frame speech signal is  $x_n(m)$ , the formula of the short-term energy is:

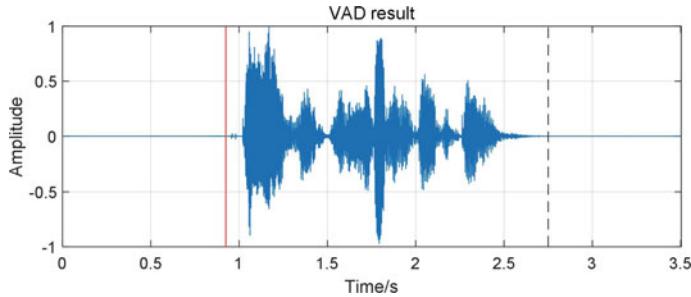
$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \quad (1)$$

where  $N$  is the frame size. The short-term zero crossing rate is the number of times that the sample changes the symbol. So the formula of the short-term zero crossing is:

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} |\text{sgn}[x_n(m)] - \text{sgn}[x_n(m-1)]| \quad (2)$$

where  $\text{sgn}[x]$  is a symbolic function:

$$\text{sgn}[x] = \begin{cases} 1, & (x \geq 0) \\ -1, & (x < 0) \end{cases} \quad (3)$$



**Fig. 2** The result of double thresholds VAD

The short time energy detection can distinguish the voiced and unvoiced speech accurately. But unvoiced speech can be mistakenly judged as silence because they have very low short-term energy. The short time zero crossing rate can distinguish the silence and the unvoiced speech effectively. Thus in double thresholds VAD, the combination of the two parameters, the speech segment and silent segment can be detected accurately [14].

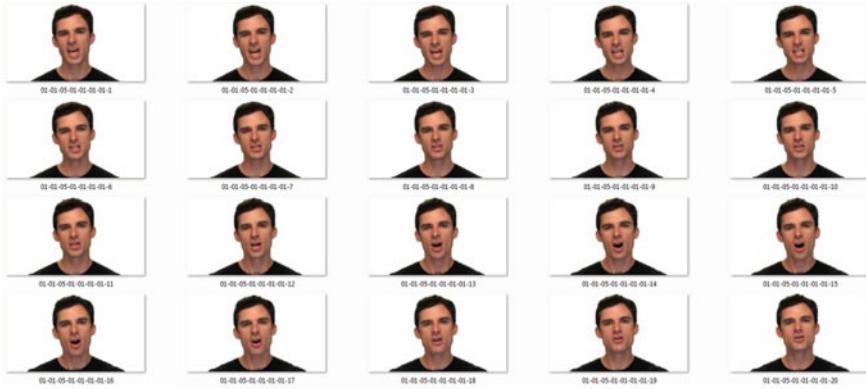
In this paper, two thresholds of the short-time energy and a threshold of short-time zero crossing rate are set with empirical values. In the voice activity detection, we also define the maximum pause time and the minimum speech segment length. If the detected silence segment does not exceed the maximum pause time, it is considered that the speech segment hasn't ended. If the detected speech segment is less than the minimum speech segment length, it is considered that the speech segment is invalid and will be modified to silence segment.

In Fig. 2, the position of the red solid line and the black dotted line represents the starting point and the end point of the speech segment respectively. The speech segment begins at the 37th frame and ends at the 110th frame, and the other 74 frames are considered as silence segments according to the VAD result.

### 2.3 *Capturing and Refining of Facial Expression Images*

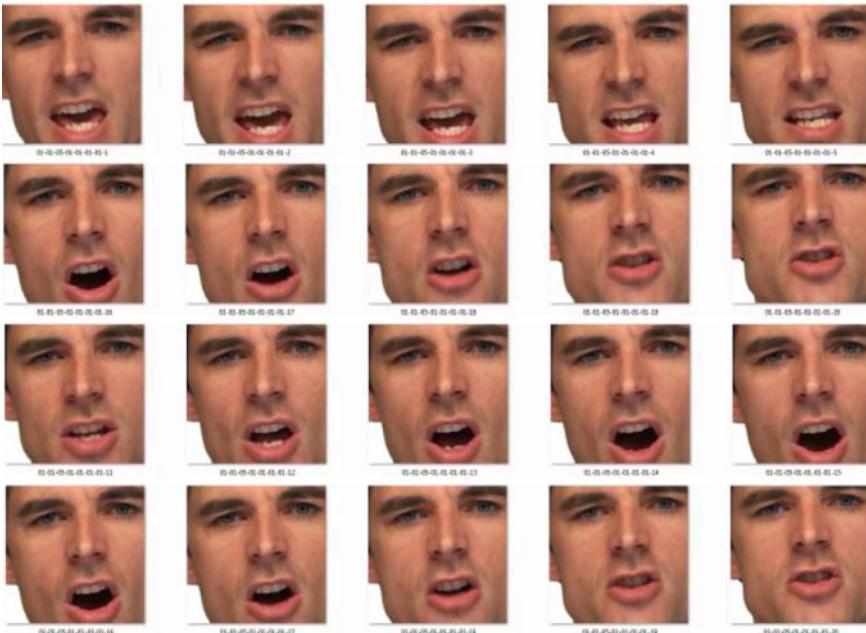
As mentioned above, there are the emotional audio and the facial expressions images in the emotional video files, and facial expression is more obvious and stronger during the speech period. We can find the endpoints of the audio component in the video using voice activity detection in the audio modal to capture images with better emotional recognizability. VAD for the audio modal is an important refined procedure in the capturing of facial expression from video files.

Some images captured according to the results of VAD are shown in the Fig. 3. It's clear that the images express intensive emotion and there aren't images in the silence segments that is hard to recognize emotion as in Fig. 1.

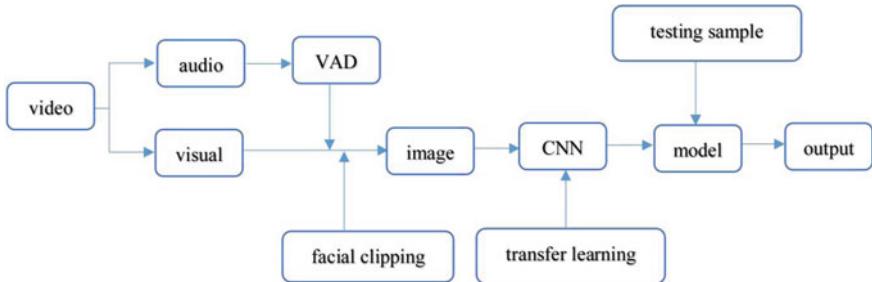


**Fig. 3** Facial expression images captured according to the VAD result

Facial clipping is implemented after the image capturing. In this procedure, compared to the images shown in Fig. 3 the useless part of the images such the background and the speaker's body except for the face will be clipped. The facial expression images are clipped to the size of  $224 \times 224$ . Some clipped images are shown in Fig. 4



**Fig. 4** Facial clipping after the images capturing



**Fig. 5** The research diagram of this paper

### 3 The Research Diagram of Facial Expression Recognition

Combined with the discussion above, Fig. 5 illustrates the research diagram of this paper.

For the emotional video files, the audio and visual modals are synchronous and can be processed separately. Double thresholds VAD is performed on the audio signal. According to the VAD results, the images synchronous with the speech segments are captured and the images synchronous with the silence segments are discarded. After the facial clipping of the captured image, the facial expression images with better emotion recognizability are obtained as the data set. In the learning and training of facial expression recognition model, this paper carries out transfer learning on the existing pretrained CNN model. And finally, the testing images can be recognized by the model and output the recognition results.

In this paper, pretrained ImageNet-VGG-F model is adopted. VGG is a CNN architecture. It was proposed by Karen Simonyan and Andrew Zisserman of Oxford Robotics Institute in 2014. It was submitted to Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) and the model achieves 92.7% top-5 test accuracy in ImageNet, which is one the largest data-set available with 14 million hand-annotated images. VGG was a breakthrough of Convolutional Neural Networks after LeNet-5 (1998), AlexNet (2012), ZFNet (2013), and GoogleNet/Inception (2014). This Fast (CNN-F) architecture comprises 8 learnable layers, 5 of which are convolutional, and the last 3 are fully connected [15]. The input image size is  $224 \times 224$ .

### 4 Experiments and Results

In this paper, facial images of speaker 01 in the RAVDESS dataset are studied. In the RAVDESS dataset, emotions are labeled as 1 = neutral, 2 = calm, 3 = happy, 4 = sad, 5 = angry, 6 = fearful, 7 = disgust, and 8 = surprised. Emotional intensity has normal and strong and there is no strong intensity for the ‘neutral’ emotion. Totally 60 video files including audio-visual modals are used in experiments.

The experiments adopt MatConvNet toolbox which is a MATLAB toolbox implementing Convolutional Neural Networks (CNNs) [16]. It is simple, efficient, and can run and learn state-of-the-art CNNs. Many pretrained CNNs for image classification, segmentation, face recognition, and text detection are available. The pretrained ImageNet-VGG-F model is adopted in the facial expression recognition. After captured with the results of VAD and image clipping, all images are resized to  $224 \times 224$ . In the learning procedure, the epoch time is 200, and the learning rate is 0.001. We will compare the experimental results with and without VAD in the image capturing of facial expression.

#### **4.1 Facial Expression Recognition Results Without VAD**

For the speaker01 in the RAVDESS database, all the images will be captured in the visual modal if VAD algorithm isn't adopted to the audio modal in the video file. 20 images are selected randomly from every video as testing images, so there are 1200 testing facial expression images for 60 videos of eight emotions. And the rest 5543 images of eight emotions are the training images. In this case, the facial expression recognition accuracy is 72.167%, and the corresponding confusion matrix is listed in Table 1.

In the confusion matrix, the labels 1–8 correspond to the emotion types mentioned above. The data at diagonal position represents the correct recognition accuracy of

**Table 1** Confusion matrix without VAD (%)

Labels	1	2	3	4	5	6	7	8
1	<b>47.500</b>	3.750	13.750	0.000	0.000	0.000	0.000	35.000
2	0.000	<b>36.250</b>	63.125	0.000	0.000	0.000	0.000	0.625
3	0.000	0.000	<b>100.000</b>	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	<b>100.000</b>	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	<b>55.000</b>	1.250	3.750	40.000
6	0.000	0.000	25.000	3.125	0.000	<b>26.875</b>	2.500	42.500
7	0.000	0.000	0.000	0.625	0.000	0.000	<b>99.375</b>	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	<b>100.000</b>

Source: Bold values represent the correct emotion recognition accuracy of facial expression images

facial expression, and the data at other positions indicates confusion degree between the recognized result and other emotions. The results show that there are serious confusions between emotion 1 and 8, 2 and 3, 5 and 8, 6 and 3, as well as 6 and 8. Among them, emotion 6 (fearful) has the lowest recognition accuracy and the performance on the recognition emotion 2 (calm) is also very bad. The overall experimental data shows that if all of the facial expression images in the video are captured, a large number of images without emotional recognizability are used in the learning and training procedure, which results in poor performance of the facial expression recognition model.

## 4.2 Facial Expression Recognition Results with Synchronized VAD

According to the results VAD, only the images with synchronous speech segments will be captured and the images corresponds to the silence segments will be discarded during the capturing of facial expression images. In this case, totally 4942 training images and 1200 test images of eight emotions will be obtained. The experimental results with synchronized double thresholds VAD is 91.667%, and the corresponding confusion matrix is shown in Table 2.

Compared to the experimental data in Sect. 4.1, Table 2 shows that serious confusions between emotion 2 and 3, 5 and 8 have been eliminated totally, and the confusion between emotion 1 and 8, 6 and 3, 6 and 8 have been alleviated. The emotion 2 and 5 can be recognized accurately from the facial expression images, and the recognition accuracy of emotion 6 that is fearful is 75.000%, which is much higher than 26.875% without VAD. The total recognition accuracy has been improved 19.5% higher than 72.167% in Sect. 4.1. However, the emotion 1 that is neutral is still hard to be recognized. One of the most important reasons is the number of the neutral samples. In RAVDESS, the neutral only has one intensity so this emotion has a half

**Table 2** Confusion matrix without VAD (%)

Labels	1	2	3	4	5	6	7	8
1	<b>50.000</b>	25.000	0.000	0.000	0.000	0.000	0.000	25.000
2	0.000	<b>100.000</b>	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	<b>100.000</b>	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	<b>87.500</b>	0.000	0.000	0.000	12.500
5	0.000	0.000	0.000	0.000	<b>100.000</b>	0.000	0.000	0.000
6	0.000	0.000	12.500	0.000	0.000	<b>75.000</b>	0.000	12.500
7	0.000	0.000	0.000	0.000	0.000	0.000	<b>100.000</b>	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	<b>100.000</b>

Source: Bold values represent the correct emotion recognition accuracy of facial expression images

samples of the others. There aren't adequate images captured from the neutral video in the learning procedure. In general, the facial expression recognition performance using double threshold VAD algorithm is better than the performance of the facial expression model trained by imaged without refining.

## 5 Conclusions

In this paper, we analyze the redundancy of the facial expression images captured directly from the video file, and use the double thresholds VAD results of the synchronous audio modal to refine the capturing of facial expression images. This method ensures that the images used in facial expression recognition have better emotional recognizability. A pretrained CNN model with transfer learning has been adopted in the experiments. Experimental results show that most of the serious confusion has been eliminated and the total accuracy of facial expression recognition has been improved to 91.667%.

In our future work, we will extend the scale of emotional video, especially natural emotion videos. The research can expand to nonspecific speaker, and we also can use emotional information from other modalities such as physiological signals of our human body to assist facial expression recognition.

**Acknowledgements** This work was supported by the Ph.D. Foundation Program of University of Jinan (No. XBS1929), and the Science and Technology Program of University of Jinan (No. XKY2064).

## References

1. Bänziger, T., Grandjean, D., Scherer, K.R.: Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). *Emotion* **9**(5), 691–704 (2009)
2. Alonsomartín, F., Malfaz, M., Sequeira, J., et al.: A multimodal emotion detection system during human-robot interaction. *Sensors* **13**(11), 15549–15581 (2013)
3. Poria, S., Cambria, E., Hussain, A., et al.: Towards an intelligent framework for multimodal affective data analysis. *Neural Netw.* **63**, 104–116 (2015)
4. Shiqing, Z., Shiliang, Z., Tiejun, H., et al.: Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* **28**(10), 3030–3043 (2018)
5. Hassan, M.M., Alam, M.G.R., Uddin, M.Z., et al.: Human emotion recognition using deep belief network architecture. *Inf. Fusion* **51**, 10–18 (2019)
6. Yaxiong, M., Yixue, H., Min, C., Chen, J., et al.: Audio-visual emotion fusion (AVEF): a deep efficient weighted approach. *Inf. Fusion* **46**, 184–192 (2019)
7. Rouast, P.V., Adam, M.T.P., Chiong, R.: Deep learning for human affect recognition: insights and new developments. *IEEE Trans. Affect. Comput.* (2019). <https://doi.org/10.1109/TAFFC.2018.2890471>

8. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1 (NIPS'12), pp. 1097–1105. Curran Associates Inc., Red Hook, NY, USA (2012)
9. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)
10. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning (Adaptive Computation and Machine Learning Series). MIT Press (2016)
11. Olivas, E.S., Guerrero, J., Sober, M.M., et al.: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques-2 Volumes. IGI Publishing (2009)
12. Ekman, P.: Emotions Revealed, 2nd edn. Times Books, New York (2003)
13. Livingstone, S.R., Russo, F.A.: The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **13**(5), e0196391 (2018)
14. Lawrence, R.R., Ronald, W.S.: Theory and Applications of Digital Speech Processing. Prentice-Hall Inc. (2011)
15. Chatfield, K., Simonyan, K., Vedaldi, A., et al.: Return of the Devil in the Details: Delving Deep into Convolutional Nets. Computer Science (2014)
16. Vedaldi, A., Lenc, K.: MatConvNet—Convolutional neural networks for MATLAB. In: Proceedings of Conference on Multimedia. ACM (2015)

# An Unsupervised Concrete Crack Detection Method Based on nnU-Net



Xinyang Li<sup>ID</sup>, Shaowu Yang<sup>ID</sup>, and Hengzhu Liu<sup>ID</sup>

**Abstract** Concrete crack detection is critical to the maintenance of infrastructure. Neural network-based vision methods are widely used to address this challenging task. Current supervised learning methods rely heavily on a large amount of labeled data. To tackle this problem, we proposed an unsupervised concrete crack detection method based on nnU-Net, in which a new reconstruction strategy is applied. Specifically, the input image is first passed to the trained model, and the output image is used as the input image of the final model to reconstruct the original image. We introduced the weighted sum of  $L_2$  and SSIM as the loss function of the network to improve the reconstruction ability of the model. We analyzed the proposed method and compared it with the baseline method, which shows obvious improvement. The method has an AP of 0.84 and AUROC of 0.85 on the test set and can give rough estimate of concrete crack location.

**Keywords** Crack detection · Reconstruction strategy · nnU-Net

## 1 Introduction

Concrete infrastructure is an important part of our living environment, which includes bridge decks, walls, and pavements etc. Due to the effect of weather, man-made and other factors, the aging and damage of concrete infrastructure has become an inevitable problem. For the safety of people's lives and property, it is necessary to detect the concrete crack early and effectively, which can help authorities to make the next maintenance plans and can also avoid the occurrence of malignant accidents possibly leading to casualties. Therefore, the automatic concrete crack detection has become a research hot spot.

In the past, the abnormal information of concrete is mainly obtained by human observation. Recently, structural health monitoring technology has been proposed by many researchers. Compared with the traditional concrete crack detection methods,

---

X. Li · S. Yang (✉) · H. Liu

State Key Laboratory of High Performance Computing, College of Computer, National University of Defense Technology, Changsha 410073, China

e-mail: [shaowu.yang@nudt.edu.cn](mailto:shaowu.yang@nudt.edu.cn)

it reduces lots of manpower and financial costs, makes up for the unreliability of human inspection, and also reduces the occurrence of accidents during inspection due to complex conditions. However, this kind of technology is limited for some large-scale infrastructure detection for the difficulties to deploy the detection equipment due to the complexity of the environment. In recent years, some computer vision methods based on neural network have been applied in many research fields, such as object detection, object tracking, image segmentation, which have achieved promising results. How to exploit the computer vision methods to detect the concrete crack automatically with the least amount of manpower becomes a hot spot. Many researches transform the problem of concrete crack detection into target detection and image segmentation task. In these tasks, a large number of labeled samples are used to train the model, so as to judge whether there are anomalies in the test concrete images and give abnormal regions by the model. However, with the increasing of labor costs, annotated data sets are often not available. Therefore, using unsupervised learning method to solve the problem of concrete crack detection has attracted more and more attention.

U-Net [1] for segmentation has been used to solve challenges in multidisciplinary fields, such as biomedical image segmentation and automatic driving. In this study, an unsupervised concrete crack detection method is proposed based on nnU-Net [2] which is an adaptive network based on U-Net. We formulate the challenge as an anomaly detection task that can be solved using a reconstruction strategy. The concrete images with high reconstruction losses are most likely to be the abnormalities. Thus, we adopt U-Net architecture to reconstruct the concrete image. In order to learn the feature representation of normal samples better, a training strategy is proposed, in which the blurred image generated by the previously trained model is used as the input of the final model. Furthermore, we optimize the model by making the output as similar as possible to the original clear image. We first train the model using the proposed method on the public dataset and then evaluate the algorithm's performance at the sample level on the test set. We introduce two common evaluation metrics to evaluate the model: Average Precision (AP), Area under Receiver Operating Characteristic (AUROC). The performance of the proposed method is examined by comparative studies. It can be proved that the performance of this method is better than the method based on the original image reconstruction by horizontal contrast experiment, and it is also better than other artificial reconstruction strategies by longitudinal contrast experiment.

## 2 Related Work

Currently, computer vision methods have been applied in concrete crack detection, achieving promising results. Such methods can be divided into two classes. The first category focuses on traditional image processing technology, and the second category is devoted to use the deep convolutional neural network to deal with the concrete crack detection issue.

For those methods based on image processing technology, Dinh et al. [3] addressed the problem of concrete crack detection by handling histogram thresholding to collect regions of interests from the background. Nhat-Duc [4] proposed a method for adjusting gray intensity, which can improve the accuracy of the detection results of the crack. Even though the methods based on image processing technology have the advantage that almost all surface defects can be identifiable, they do not perform well in crack detection tasks under complex backgrounds because they are difficult to extract high-level features. Besides, a suitable classifier algorithm needs to be designed in this kind of method which is also a challenging task.

In recent years, methods based on deep learning have shown a promising prospect in the field of computer vision. Deep learning technology is a kind of data-driven method. In its process, a specific model structure is selected by the characteristics of data sets and tasks, and the optimization algorithm is applied to adjust the model parameters. There are three main approaches for the concrete crack detection by deep learning technology which are boundary box regression method, image classification method based on patch and semantic segmentation method. They almost need a large quantity of labeled data. These methods are briefly described in the following part.

The first method is based on object detection approach. The object detection technology regresses the bounding box coordinates to train model. Chen et al. [5] proposed the NB-CNN to detect individual video frames using a Naïve Bayes data fusion scheme and a convolutional neural network. Cha et al. [6] proposed a structural visual inspection method based on Faster R-CNN. This kind of method usually gives a rectangular box to mark the final crack position. Due to the irregularity of the crack, crack patterns tend to occupy only a small proportion in the bounding box. Therefore, it can only detect the crack position roughly. The second method is based on image classification combining with the sliding window technique. A classification model is trained and consequently, it is used to classify the patches divided from the entire image by different window sliding strategies. Typical example is the approach proposed by Cha et al. [7]. The above two methods can rarely detect concrete cracks in pixel level. To tackle this challenge, some researchers treat the concrete crack detection task as a semantic segmentation problem. The method classifies every pixel in the concrete inspection into crack and non-crack pixel. Fully convolution network (FCN) [8] is commonly used in this method. Dung et al. [9] presented a concrete crack detection method based on FCN. They firstly chose the backbone of FCN encoder by the image classification performance of the network on a public concrete crack dataset. Then, the entire FCN network was trained end-to-end on an annotated dataset for semantic segmentation. Based on the success of U-Net in semantic segmentation for biomedical images, Liu et al. [10] combined the U-Net and the FCN structure to detect concrete damage with smaller training set but with good accuracy.

In practice, annotating the ground truth to obtain the labeled dataset by hand would be prohibitively expensive and impractical. In recent years, detection of cracks using unsupervised learning methods is a crucial task. Mubashshira et al. [11] presented an unsupervised approach based on K-means clustering algorithm and Otsu thresholding for road surface crack detection. With the proposal of Generative Adversarial

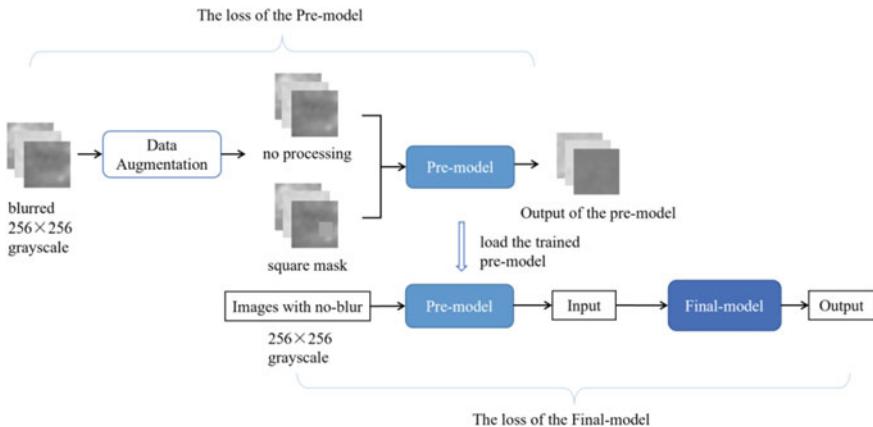
Network (GAN) [12] in 2014, GAN-based research has become a research hot spot, GAN has also been applied to many fields such as domain adaptation [13], image in painting [14] and biomedical anomaly detection [15]. Duan et al. [16] present an unsupervised pavement crack detection method based on GAN to learn the mapping from crack images to binary images with unpaired data. Li [17] exploits GAN to make encoder-decoder network generate reconstructed images better by means of adversarial learning. Although the unsupervised learning method based on GAN can detect the concrete crack relatively accurately, its model construction is complicated and the training is difficult to converge. Considering the problems mentioned above, we propose an unsupervised concrete detection method based on nnU-Net. Generally, our contributions are as follows:

- (1) We model the concrete crack detection problem as an anomaly detection task which can be solved using a reconstruction strategy. The concrete images with high reconstruction losses are most likely to be the abnormalities.
- (2) Compared with the scheme of adding manual rules such as occlusion and rotation to the original image as the input image of the encoder-decoder, this method adopts the means of automatic model generation and avoids the step of artificially adjusting super parameters to adapt to the model.
- (3) In the process of training, we use the loss function formed by weighting L2 loss and structural similarity loss (SSIM). In order to make up for the fact that L2 distance is unable to measure the structural similarity of images, we introduce SSIM loss and adjust the weighting ratio by the validation set.

### 3 Proposed Method

#### 3.1 Overall Procedure

The proposed method consists of two encoder-decoder networks which are termed as pre-model and final-model. The raw concrete images are preprocessed first to reduce the effect of illumination and messy background. We convert the images into grayscale mode from RGB mode. Then, the images are averaged filtered and the convolution kernel's size is set to be 25. Subsequently, we train the pre-model whose network structure is nnU-Net using the dataset after average filter processing. A mask strategy is applied to the input image with a probability of 0.5, and the model is trained to reconstruct the image that is not applied to the mask strategy. The pre-model is saved to a file after fully training. In the second stage, the same network structure is applied to the second encoder-decoder model. In the process of the final-model training, the pre-model is loaded. The greyscale image without average filtering is sent to the pre-model to get blurred image by forward propagating which is taken as the input image of the final-model. The utility of the final-model is to make the output image of the network closer to the original greyscale image



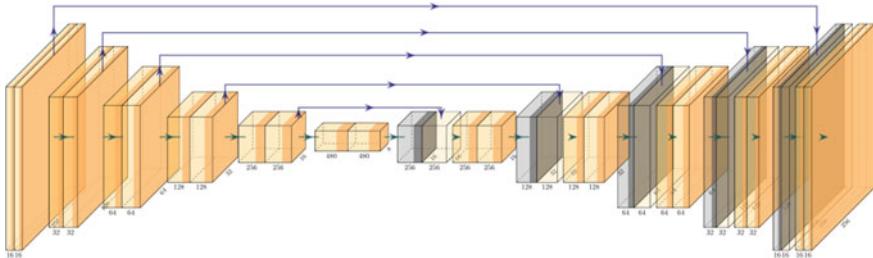
**Fig. 1** The framework of the method

without filter processing by adjusting the parameters. The overall procedure schema is described in Fig. 1.

In the process of detecting cracks, the greyscale image after average filter is passed into the final-model to get the reconstruction image. The concrete images with high reconstruction losses are most likely to be the images with crack. A post processing technique is exploited to the reconstruction images to eliminate the scattered components caused by noise through judging the amount of the component pixels.

### 3.2 Model Description

The nnU-Net is applied to both the network structure of pre-model and final-model. The framework of nnU-Net is based on the original U-Net network ensembled with an automated pipeline containing pre-processing, data augmentation and post-processing for biomedical images. The purpose of the network is to learn the characteristics of non-crack samples and reconstruct the normal images better. Thus, the network is comprised of three stages: the encoder, the decoder and the connection layers between them which can be seen as the bottleneck. An image with size  $1 \times 256 \times 256$  is input into the auto-encoder network. After the computation, the size of final output image is the same as the size of the input image, which is the reconstruction of the original image. The encoder and the decoder are composed of the convolution blocks and the convolution localization blocks respectively. The convolutional block is a collection of convolutional layers, dropout layers, batch normalization layers and the LeakyRelu layers. Ten such blocks make up the encoder network. Two of them can be regarded as a group where the size of the feature maps is the same. The first convolution layer in each group halves the height and width of the input feature



**Fig. 2** The network structure of generic nnU-Net

maps and doubles the number of channels except for the first group. Bottleneck network generates a feature map with the largest number of channels, and its feature map size is  $8 \times 8$  which is the smallest in the whole network. As for the decoder network, the skip connection is applied to the network to reconstruct clearer images. The feature map of each convolution layer of the decoder network concatenate to the corresponding down-sampling layer in the encoder network, so that the feature map of each layer can be effectively used in the subsequent calculation. The transposed convolution layers upsample the feature maps to gradually restore the images to the input image size. The structure of the network is shown in Fig. 2. The structures and functions of the middle layers will be expressed in the following sections.

The convolution operation takes advantage of the sliding of the convolution kernel when calculating the weighted sum to share the weights of the network, which can cut down the amount of the network parameters in the case of extracting semantic features of images. In the convolution operation of the network structure, the kernel size is selected to be  $3 \times 3$  and the padding size is 1 in each of height and weight in the feature maps. Except for the first group, the convolutional stride size is 2 in the first block of each group to achieve downsampling. In order to reconstruct the input data, upsampling operation is essential. In this structure, transposed convolution layers are used in the decoder. The convolution stride is 2. In this paper, the dropout layers zero out each channel independently on every forward call with probability 0.5 using samples from a Bernoulli distribution. Batch normalization [18] is always added after the convolutional layer of the convolution neural network for data normalization processing. The mathematical representation is as follows. The input size for batch normalization layer is 4 dimensions, which can be represented as  $(N, C, H, W)$  and the input batch data is represented by  $x$ .  $y$  stands for the output data through the batch normalization layer. The mean and standard-deviation are calculated per-dimension over the mini-batches, which are indicated by  $E[x]$  and  $Var[x]$  respectively. It can be described as

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta \quad (1)$$

The scaling factor  $\gamma$  and the translation coefficient  $\beta$  are learnable parameter vectors of the input size. The parameter  $\epsilon$  is a value added to the denominator for the purpose of numerical stability.

The activation function is used to add nonlinear factors to improve the expressive ability of linear model. One of the nnU-Net's major modifications to the network structure is to change the nonlinear layer from the original *Relu* function to the *LeakyRelu* [19] function. The *LeakyRelu* activation function is expressed as

$$\text{LeakyRelu}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \text{negative\_slope} \times x, & \text{otherwise} \end{cases} \quad (2)$$

It assigns a non-zero slope to all negative values instead of zero. In this paper, the negative\_slope is set to 0.01 by default.

### 3.3 Loss Function

In order to optimize the reconstruction model, we need to select a loss function. For solving the reconstruction problem, the Mean Square Error (*MSE*) loss function is commonly used to make the reconstruction image of the model as similar as possible to the input image of the model. The *MSE* loss function is expressed as

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (3)$$

Among them, I and K respectively stand for two images, and m, n represent the length and width dimensions respectively.

When there are outliers for the images, they will account for the main components of loss. It makes the gradient fluctuate greatly, which makes it difficult for the model to converge and reduces the performance of the model. The Structural Similarity (*SIMM*) [20] loss function pays more attention to the regional features of image blocks, which makes up for the deficiency of *MSE* loss function, which concerns the reconstruction loss of a single pixel, and makes the trained model more robust. The *SIMM* loss compares the similarity of two images x and y in three dimensions: luminance  $l(x, y)$ , contrast  $c(x, y)$  and structure  $s(x, y)$ , which is described as

$$\text{SIMM}(x, y) = f(l(x, y), c(x, y), s(x, y)) \quad (4)$$

The *SIMM* loss is a function of the three parts. They are defined as

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (5)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (6)$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (7)$$

In the equations,  $\mu_x$  and  $\mu_y$  are the average value of all pixels in the images' block,  $\sigma_x$ ,  $\sigma_y$  are the variances of the pixel intensity of the images and  $\sigma_{xy}$  is the covariance.

The use of  $C_1$ ,  $C_2$ ,  $C_3$  in the formula are just in case the denominator is zero. If  $C_3 = C_2/2$ , the final formula of *SIMM* is

$$\text{SIMM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (8)$$

by reduction.

The *SIMM* calculates the similarity of an image window (for example,  $11 \times 11$ ), and then averages the similarity of all windows as the similarity of the whole image. For the two images, their *SIMM* parameter is always less than 1 and 1 means exactly similar between the images. In this paper, we set the window size to  $11 \times 11$  by default. The final loss function is the weighted sum of *SIMM* loss and *MSE* loss, and the weights are adjusted by the model's performance on the validation set, so the complete loss function is

$$\text{Loss}(x, y) = 0.7 \text{ SIMM}(x, y) + 0.3 \text{ MSE}(x, y) \quad (9)$$

## 4 Experiments

### 4.1 Implementation

All the tasks expressed in this paper were implemented on a server cluster. We used two TITIAN V graphics cards with 32 GB display memory in the cluster for the experiment. The configuration for the CPU was IntelR XeonR E5-2620v4 @2.10 GHz. The operating system was Ubuntu 16.04.6 LTS. We used pytorch1.5.0 as the framework for deep learning.

We used Adam optimizer with learning rate 0.0001 to adjust the model parameters. The HE initialization method [21] is applied to the model for weights and bias initialization. We set the epochs of training to 25 and the batch size to 16. The super parameters used in both of the pre-model and final-model are the same.

## 4.2 Datasets

An open source dataset for concrete crack classification [22] is applied in our experiment. The pictures in the dataset are obtained from METU Campus buildings. The dataset is divided into two categories as non-crack and crack images, that each has 20,000 items of  $227 \times 227$  pixels, which have RGB channels. They are generated from 458 complete pictures of  $4032 \times 3024$  pixels through the method proposed by Zhang et al. [23]. The sample images having no crack on concrete are indicated in Fig. 3. The sample images with crack on concrete are indicated in Fig. 4.



**Fig. 3** Sample concrete surface images with no-crack



**Fig. 4** Sample concrete surface images with crack

The training set for the concrete crack detection task contain 18,000 normal images. The validation and test sets both contain 2000 images, half of which are normal samples and half are samples with cracks. In order to meet the input size requirements of the neural network, we use bilinear interpolation to resize the image from  $227 \times 227$  to  $256 \times 256$ . Then we converted the RGB images into greyscale images whose the number of channel is 1. Finally, we applied the maximum and minimum normalization method to normalize the image data to make the network training more reasonable.

### 4.3 Evaluation

We take the difference between an image and its reconstructed image pixel by pixel to calculate the square, and the resulting image is called the difference image. The average value of the intensities in the difference image is regarded as an abnormal score for the image. As the noise often exists on the background of concrete surface, it brings the reconstruction errors, which will affect the judgment of the real crack area we are concerned about and make it difficult to distinguish between normal image and abnormal image. Therefore, we further process the difference image by giving the threshold of the number of pixels in the connected domain of the image, according to the fact that the pixels in the real crack area is well connected, while the connected pixels that constitute the noise is not. We reduce the pixel value of the possible noise area by multiplying the pixel value of the area with a small connected number by a decimal number to reduce its proportion in the average value result.

The AP and AUROC are the comprehensive metrics that evaluate the resolution ability of the model. The precision (Pr) and recall (Re) are the basic measures of the capabilities of the classification system, and are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

where FP, TP, FN are the numbers of false positive, true positive and false negative respectively. Given different classification thresholds, a set of values of Pr and Re will be obtained, which could consist of coordinates. The area enclosed by the curve is the AP value. The larger the AP value is, the better the effect of the classifier. With FPR as the horizontal axis and TPR as the vertical axis, the area enclosed by the curve and the coordinate axis has turned into another commonly used metric for measuring the performance of the classifier, which is AUROC. It is an excellent metric for studying the generalization performance of the learner. We select these two comprehensive metrics AP and AUROC to evaluate the performance of the model.

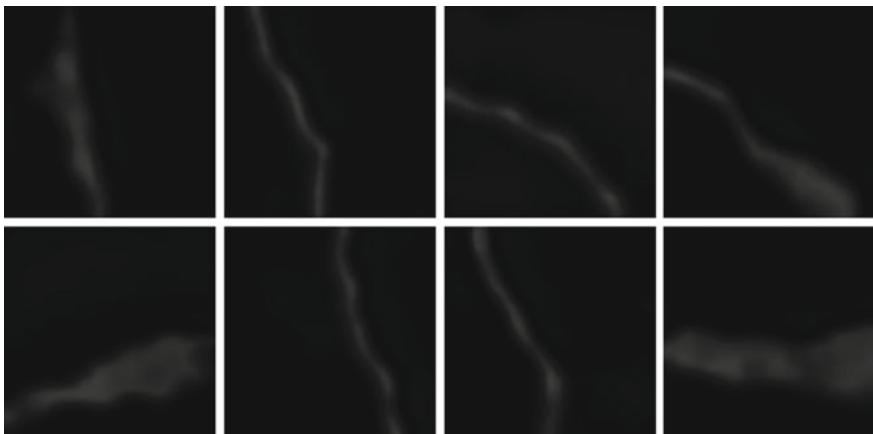
#### 4.4 Results and Discussions

We compared our model with other auto-encoder models, including convolutional auto-encoder (CAE) [24], U-Net [1] and U-Net3+ [25]. The trained model was evaluated on the test set, where both of the amount of positive and negative samples are 1000. The AP and AUROC calculated on test set are shown in Table 1. It can be identified that the performance of our proposed method is better than that of other generation models in the table. Figures 5 and 6 shows the difference images for crack images and non-crack images respectively, in which the locations of cracks are indicated.

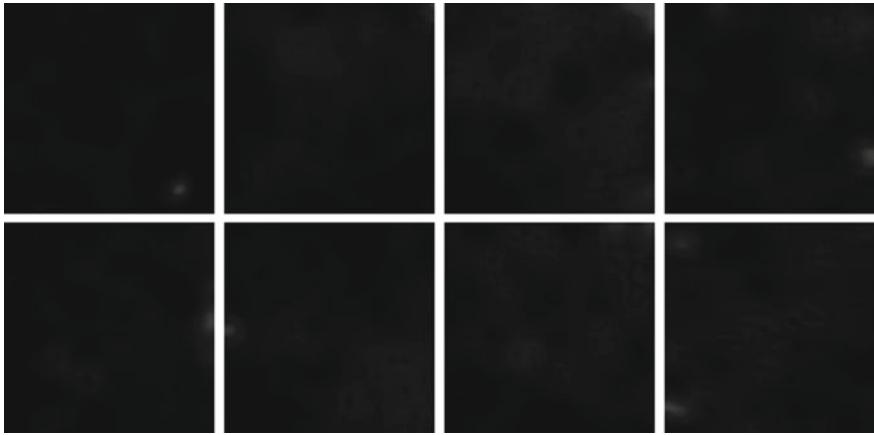
In order to examine the superiority of the model-based method of automatically generating self-encoder input images, we designed the comparative experiments, including horizontal and longitudinal comparisons. As for that horizontal contrast experiment, we directly use the original image that has not been processed by the model as the input image of the auto-encoder. We compare it with the model that uses the image obtained from the output of the pre-trained model as the input image of the final model, in which the structures of models are both nnU-Net expressed in the above section. The experimental results are shown in the Table 2.

**Table 1** Comparison of different methods

Method	AP	AUROC
CAE	0.348	0.220
U-Net	0.399	0.274
U-Net3+	0.477	0.405
The proposed method	0.840	0.853



**Fig. 5** Difference image samples with crack



**Fig. 6** Difference image samples with non-crack

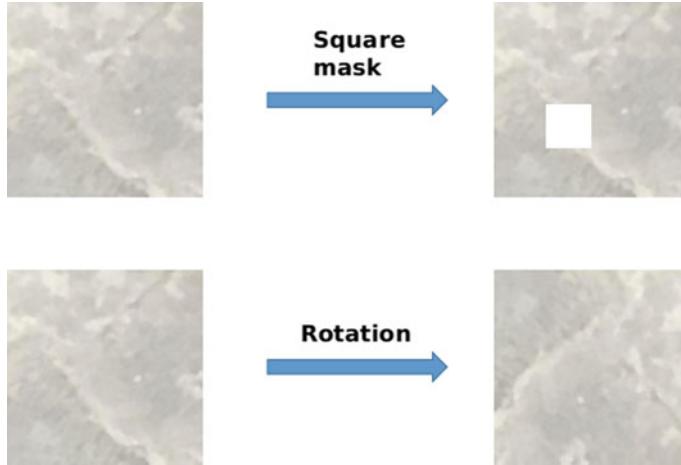
**Table 2** The results of horizontal comparison experiments

Method	AP	AUROC
The original reconstruction strategy	0.457	0.261
The reconstruction strategy based on model generation	0.655	0.527

Through the training of previously mentioned two models, the second model is used as supplement to the first model. The input image is first reconstructed by the first model to obtain a relatively fuzzy reconstructed image. The second model performs secondary reconstruction on the input blurred image, reconstructing the area that the first model failed to reconstruct to achieve stronger feature reconstruction ability. As a result, the normal sample is better reconstructed, i.e., the reconstruction error of the image is smaller. Thus, the image with crack and the image with non-crack can be better distinguished.

Some other works [26] use artificially chosen masking block to cover the images, and train the network to fill in the missing areas with the contextual information of the pictures, thereby promoting the reconstruction performance of the auto-encoder. Although this method is relatively simple to implement, it needs to manually adjust the parameters of the masking block, such as the size of the masking block, and the pixel value of the filling. The rotation-based reconstruction strategy enables the network to reconstruct the image before rotation. It promotes the generalization performance of the network. The schematic diagrams of these two strategies are as follows (Fig. 7).

The method we propose uses the network to generate the input image of the auto-encoder without artificial adjustment of the corresponding parameters. Table 3 indicates the results of longitudinal comparison experiments. As we can identify, the performance of using the method with automatically generating the input images



**Fig. 7** The transformations to the concrete image

**Table 3** The results of longitudinal comparison experiments

Reconstruction strategy	AP	AUROC
Rotation	0.321	0.105
Mask	0.490	0.312
Ours	0.655	0.527

through the network is better than that of using the artificially formulated strategy for transforming the original input images.

## 5 Conclusion

The reconstruction-based concrete crack detection method is more suitable for the case where the variance of the normal sample distribution is small. In this case, the normal image can be better reconstructed by fully learning the semantic information of the normal sample, while abnormal image cannot. The abnormal area is detected by the reconstruction error. The increase in reconstruction errors of normal images makes it more difficult to figure out between crack and non-crack images. Therefore, we preprocess the images by blurring to reduce the effect of noises, so that the non-crack images can be better reconstructed. Experiments show that it is feasible to transform the problem of concrete crack detection into an anomaly detection task, which can detect concrete cracks at sample level and give the crack location roughly without labels. The proposed reconstruction strategy and the choice of loss function promote the network to reconstruct the image better. We have made some attempts in using unsupervised methods to solve the concrete crack detection problem. Because

the image is blurred and there is no auxiliary supervision information such as segmentation mask, this method has limitations in pixel-level crack detection. Our future work will explore methods that can improve the accuracy of pixel-level detection, with robust models which are more inclusive of image noises.

**Acknowledgements** This work was partially supported by the National Natural Science Foundation of China (No. 61803375 and 91948303), and the National Key Research and Program of China (No. 2017YFB1001900 and 2017YFB1301104).

## References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv.abs/1505.04597 (2015)
2. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S.A., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S.J., Maier-Hein, K.: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv.abs/1809.10486 (2018)
3. Dinh, T., Ha, Q., La, H.: Computer vision-based method for concrete crack detection. In: Proceedings of the International Conference on Control, Automation, Robotics and Vision, pp. 13–15. Phuket, Thailand, November (2016)
4. Hoang, N.: Detection of surface crack in building structures using image processing technique with an improved otsu method for image thresholding. Adv. Civ. Eng. 1–10 (2018)
5. Chen, F., Fahanshahi, M.: NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion. IEEE Trans. Ind. Electron. **65**, 4392–4400 (2018)
6. Cha, Y., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O.: Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Comput.-Aided Civ. Infrastruct. Eng. **33**, 731–747 (2018)
7. Cha, Y., Choi, W., Büyüköztürk, O.: Deep learning-based crack damage detection using convolutional neural networks. Comput. Aided Civ. Infrastruct. Eng. **32**, 361–378 (2017)
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
9. Dung, C.V., Anh, L.: Autonomous concrete crack detection using deep fully convolutional neural network. Autom. Constr. **99**, 52–58 (2019)
10. Liu, Z., Cao, Y., Wang, Y., Wang, W.: Computer vision-based concrete crack detection using U-net fully convolutional networks. Autom. Constr. **104**, 129–139 (2019)
11. Mubashshira, S., Azam, M.R., Ahsan, S.M.: An unsupervised approach for road surface crack detection. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1596–1599 (2020)
12. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
13. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 95–104 (2017)
14. Demir, U., Ünal, G.B.: Patch-Based Image Inpainting with Generative Adversarial Networks. arXiv.abs/1803.07422 (2018)
15. Zhang, C., Wang, Y., Zhao, X., Guo, Y., Xie, G., Lv, C., Lv, B.: Memory-augmented anomaly generative adversarial network for retinal OCT images screening. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1971–1974 (2020)
16. Duan, L., Geng, H., Pang, J., Zeng, J.: Unsupervised pixel-level crack detection based on generative adversarial network. In: Proceedings of the 2020 5th International Conference on Multimedia Systems and Signal Processing (2020)

17. Li, Q.: A concrete crack recognition method based-on autoencoder. *J. Beijing Jiaotong Univ.* **44**(2), 98–104 (2020)
18. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv.abs/1502.03167* (2015)
19. Maas, A.L. Rectifier Nonlinearities Improve Neural Network Acoustic Models (2013)
20. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004)
21. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034 (2015)
22. Åzgenel, Å., SorguÅ, A.G.: Performance Comparison of Pretrained Convolutional Neural Networks on Crack Detection in Buildings (2018)
23. Zhang, L., Yang, F., Zhang, Y., Zhu, Y.: Road crack detection using deep convolutional neural network. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3708–3712 (2016)
24. Masci, J., Meier, U., Ciresan, D., Schmidhuber, J.: Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. *ICANN* (2011)
25. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y., Wu, J.: UNet 3+: A full-scale connected UNet for medical image segmentation. In: ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059 (2020)
26. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.: Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. *arXiv.abs/1812.05941* (2018)

# Robust Facial Landmark Localization Based on Texture and Pose Correlated Initialization



Junwei Zhou , Mengying Li , and Yiyun Pan

**Abstract** Recently, cascaded pose regression has attracted more and more attention because of its superior performance in facial sign localization and occlusion. However, this method is very sensitive to initialization, in which incorrect initialization will seriously reduce performance. In this paper, we propose cascaded pose regression (rcpr). By checking the local binary pattern histogram of the test face and the face in the training data set, the most related to the test face is selected initialization. In order to make initialization more robust to various poses, we estimate the rough pose of the test face according to the roughly. Then, the pose related initial shape is constructed from the average face shape and rough test face pose. Finally, texture dependent and pose shapes are connected together as robust initialization. We evaluated rcpr on a challenging cofw dataset. Experimental results show that the proposed scheme has better performance than the most advanced methods in facial sign localization and occlusion detection.

**Keywords** Facial landmark localization · Cascaded pose regression · Robust initialization · Occlusion · Texture and pose correlated

## 1 Introduction

Facial landmark localization, which is localizing the facial key points (e.g., eye brows, eyes, nose, mouth and jaw), detection [1, 2], face [3–5] and expression analysis [6–8]. In recent years, and even on datasets collected in the wild [9–14]. Faces with various variations in appearance.

Estimate facial shapes [15], approaches for facial landmark localization [10, 12, 16–25]. CPR and its variations. Based on CPR, Burgos proposed a scheme of Robust CPR (RCPR) [12], which is the first scheme explicitly detect occlusion state locations of landmarks. And they created (COFW) [12]. Researchers have used this dataset to study facial landmark localization under occlusions [12, 18, 21, 26–28].

---

J. Zhou · M. Li (✉) · Y. Pan

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, P.R. China

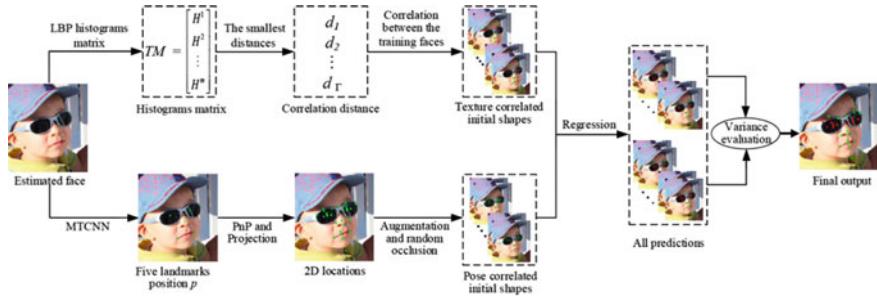
e-mail: [limengyingmmz@whut.edu.cn](mailto:limengyingmmz@whut.edu.cn)

Although these methods make some progress on facial landmark localization under partial occlusion, the occlusion problem is not essentially solved. The accuracy of occlusion prediction is still unsatisfactory. Since the occluded landmarks can hardly provide information for further analysis, it is significant to detect occlusion state of landmarks, the occluded landmarks may reduce the accuracy of localizing the un-occluded landmarks.

In recent years, based on the robust cascade regression algorithm, a new two-level cascade regression model has been proposed. This algorithm can learn more accurately the facial feature information, facial feature point positioning and the cascaded pose regression algorithm. Occlusion prediction is more accurate, which is of great significance in the field of image recognition. The two-level cascade regression method is an algorithm that combines global regression and local regression. After obtaining the face shape in the global stage, it is divided into four parts: left eye, right eye, nose, and mouth, and then performs local division and then integration, and finally get a complete face prediction picture. However, when the face contains partial occlusion, the accuracy of face feature point location is greatly limited. Because the occlusion part of the face image has too much influence on the face prediction, the performance of the regression is reduced, and the face feature point location is accurate. The rate is reduced, and even the positioning fails. In order to solve the adverse effects of occlusion on the face feature location, this paper proposes the face feature point location based on self-enhanced cascaded regression. In the training phase, the method first predicts the shape of the face initially, and calculates according to the area divided by the face. The face prediction accuracy rate of each area is calculated by calculating different weights to update the regressor according to the occlusion accuracy rate of different areas, so that the regressor can better learn the face features. In the test phase, the accuracy of the occlusion detection in each region of the face obtained by training is used to improve the regression phase regressor. Experiments show that the method proposed in this paper has better results in facial feature point location and occlusion prediction.

The final output of the network is a heat map of facial feature points. According to this heat map, the spatial information of each feature point and the location coordinates of the feature point can be obtained. After testing the overall network design and exchanging different layer modules based on these annotations, switching from the standard convolutional layer with large filters and no steps to the residual learning module method, the network performance has been greatly improved. Thereby improving the performance of the facial feature point positioning algorithm. By adding heat map features, compared with traditional coordinate regression methods, the accuracy of facial feature point positioning is greatly improved.

Because regression relies on initialization, incorrect initialization will significantly reduce performance. When the face changes and the occlusion fails, the positioning will be locked. It usually leads to the failure of landmark positioning prediction. A robust initialization method is proposed, which avoids the nicks and roughness of the test surface, and obtains a clear correlation and corresponding initial shape. On the test surface and on the surface. Through the roughness and shape of the five key points, the contour of the face is obtained. Then 29 facial key points are



**Fig. 1** The procedure of RICPR. The texture correlated initial shapes and the pose correlated initial shapes are calculated in parallel. The text correlation of testing and the training faces, while the pose correlated initialization is based on the evaluated rough face pose. These initial shapes are combined together as robust initializations for regression to get predictions. Finally, the reliability of each prediction is evaluated by variance to get the final output

used to represent another 3D average facial shape. The fuzzy correlation shows that we can detect the initial shape of the face. The projection angle of the initial shape related to the rotation is larger. We first use a cascaded network (CNN) to estimate points, including pupils, nose and mouth [29]. For the two-dimensional position of the object [30–35], the corresponding initial posture related to reality is obtained. The initial posture related to the final target and the initial posture related to the posture are used together as the initial stage, as shown in Fig. 1, which is more related to the position and the true shape of the test face in the occlusion. We use the color scheme to evaluate RICPR on the COFW dataset [12]. The NME on COFW is  $6.64 \times 10^{-2}$ , which is better than the most advanced ones. These results verify the performance of RICPR on the COFW data set. Some of the ideas presented in this paper were initially reported in Ref. [36]. In this paper, we report the full and new formulation and extensive experimental evaluation of our method. The initialization not only depends on texture correlation but also landmark localization and occlusion detection are further improved.

## 2 Related Work

Because the occlusion is very serious [12, 18, 21, 27, 37]. Burgos Artizzu et al. [12] proposed for the first time in RCPR to detect the occlusion state while estimating the landmark points, and apply the occlusion state in each iteration to obtain visually different restorations. The output of the regressor depends on the occlusion prediction result. Considering that occlusion usually affects visibility annotations, Yang et al. [18] used coordinate regression forests in multiple over-segmentation confidence values for each pixel, which is called regional predictive power (RPP). Compared with RCPR, RPP has higher positioning accuracy. Yu et al. [27] opposed the regression method (CoR) by forming consensus from the estimation of the returner specific to the

conclusion. Under the premise of satisfying a specific predefined face, the landmark position of the face is determined. Improve the accuracy of occlusion detection. Liu et al. [21] proposed a cascaded regression algorithm based on adaptive shape model (CRAM) to achieve robust facial landmark positioning. In each iteration, the method is used to estimate the occlusion level of the target, and then each landmark is weighted according to the occlusion level. In addition, in the facial landmark location and occlusion detection, the occlusion adaptive weight is used to reduce noise. Compared with other methods, the accuracy of cram's occlusion detection is 80/48.45%, and the NME is  $6.68 \times 10^{-2}$  for the localization of the COWF data set.

### 3 The Proposed Scheme

In this section, we describe in detail the proposed RICPR scheme for facial landmark localization under occlusions.

#### 3.1 Robust Initialization for Cascaded Pose Regression

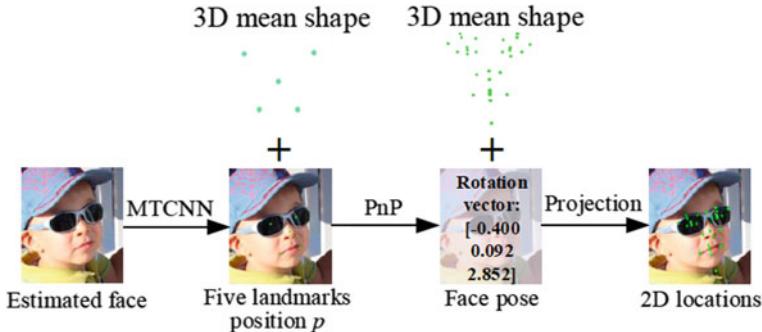
The process of the algorithm is shown in Fig. 1. First, by calculating the texture correlation between the test face and the training face, the initial shape related to the texture is obtained, and at the same time, the initial shape related to the posture is obtained by detecting the rough face pose of the test face. Then, these initial shapes are used as robust initialization for cascade regression. We respectively describe the attitude-related initialization methods in Sect. 3.2.

#### 3.2 The Pose Correlated Initial Shapes

In the above section, we describe how to select the texture correlated initial shapes considering the occlusion information but ignoring pose information of the testing face. Empirically, landmark distribution is highly correlated to head pose. To further make the initial shapes more robust to various poses, we choose some for regression.

To obtain we estimate the rough face, which can be obtained by the five landmarks, i.e., the pupils, the tip of the mouth. In this paper, we use MTCNN [29] to detect the five fiducial landmarks, as shown in Fig. 2. Inspired by Perspective- n-Point (PnP) problem, which is the problem of estimating the pose of a calibrated camera given a set of 3D points and their corresponding 2D projections in the image [38]. Given a 3D mean shape  $S$  with 5 facial key points.

Then, a 3D mean face shape, represented by 29 facial landmark locations, is projected to a set of corresponding 2D locations according to the testing face pose  $\theta^*$ , as shown in Fig. 2. After that, the shape which has similar pose with the testing



**Fig. 2** Illustration of generating the pose correlated shape. Given an image, we first detect five fiducial landmarks and estimate face pose. Then, according to the face pose, face shape with 29 facial key points, can be projected to a set of corresponding 2D locations, which has similar pose with testing image

face is obtained. To get a reasonable initial shape for each image, we re-scale the corresponding 2D locations based on the face bounding box and the detected five fiducial landmarks  $p$ .

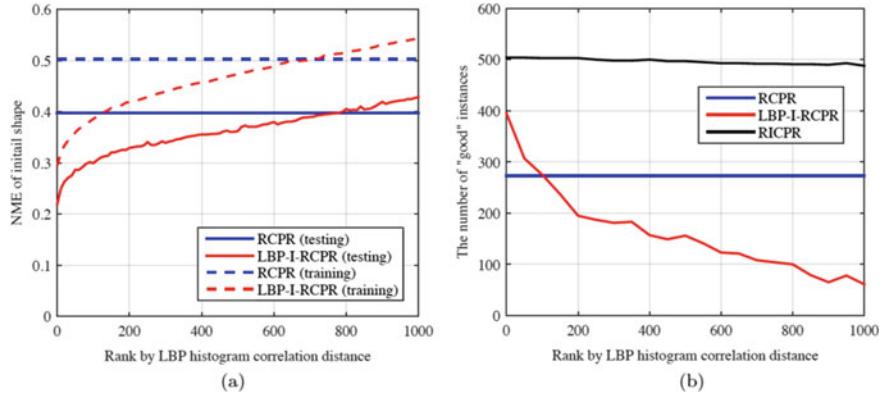
## 4 Experimental Results

### 4.1 Dataset and Implementation

The proposed scheme based on COFW data set is proposed. Face images in COFW differ greatly in shape and occlusion due to differences in postures, expressions, hairstyles, the use of accessories (such as hats) and objects (such as food, hands, microphones, etc.). Each image has 29 facial landmarks in occluded/un-occluded state. The data set has 1852 face images, of which 1345 and 507 are used for training and testing, respectively. On average, human faces account for more than 23% of COFW.

**Analysis of Initialization Based on Texture Correlation:** Instead of selecting shapes from the training set, we perform texture related initialization (I-RCPR) by calculating LBP's. In order to verify the effectiveness of the texture-related initialization method, we compared the performance of LBP-I-RCPR with the performance on the data set shown in Fig. 3.

The above NME is shown in Fig. 3a. The results show that NME decreases with the decrease of related distance, and LBP-I-RCPR can significantly reduce NME by at least 45%. It can be seen from the images based on texture correlation that this method is closer to the real shape of the human face.



**Fig. 3** Comparisons between the texture correlated initialization based RCPR and the traditional random based RCPR. **a** The NME shapes with differ and testing processes. **b** The number determined by variance after 10% cascades of each prediction

Moreover, given different initial shapes for each image, the variance between their predictions is applied to “good” class as stated in Ref. [12]. Thus less bad initial shapes are selected. Furthermore, the number of “good” instances increases from 395 to 504 among the 507 images in RICPR scheme, which means fewer than 1% instances are “bad”, thus the initialization become more robust.

We also initialize the shapes using other different features, including Local Derivative Pattern (LDP) [39], Gabor, Gaussian Markov Random Field (GMRF), Gray-Level Difference Statistics (GLDS) and Eigenface. We report the NME and occlusion detection of each feature respectively in Table 1. And performs better.

**Face Localization on COFW:** Due to the large variability of the COFW database, the performance of the positioning method on the COFW database is poor. The proposed scheme, scheme. The comparisons of NME on COFW dataset are given in Table 2.

RICPR’s NME is smallest. To get the pose correlated initial shapes, we use MTCNN to detect fiducial landmarks. The accuracy of fiducial landmarks plays a significant role on performance. If the ground-truth of the fiducial landmarks is employed, the NME can reach  $5.52 \times 10^{-2}$ , which demonstrates that the proposed scheme can obtain a admirable landmarks are detected accurately.

**Table 1** Texture correlated initialization using different features

Feature	LBP	LDP	Gabor	GMRF	GLDS	GLCM	Eigenface
NME ( $\times 10^{-2}$ )	7.35	7.75	7.87	8.28	8.19	8.06	8.18
Precision/Recall	80/51.4%	80/48.7%	80/46.1%	80/45.6%	80/47.2%	80/46.5%	80/47.6%

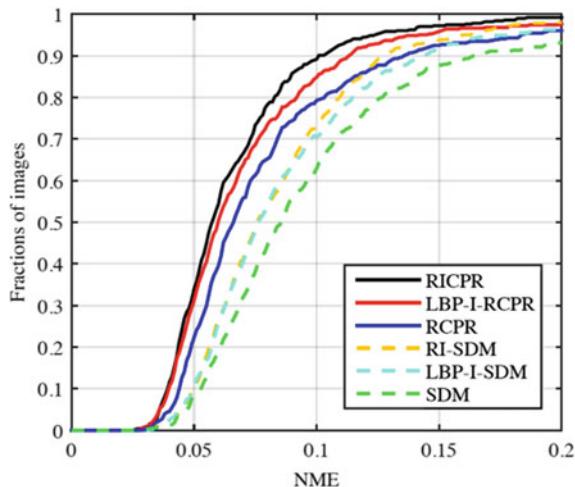
**Table 2** Results on COFW dataset

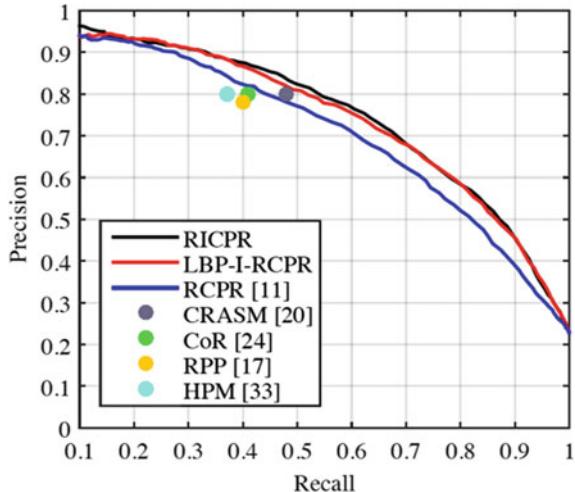
Methods	Mean error	Occlusion prediction
	NME ( $\times 10^{-2}$ )	Precision/Recall
RCPR [12]	8.01	80/42%
HPM [37]	7.46*	80/37%*
RDP [18]	7.52*	78/40%*
SDM [40]	10.88	—
TCDCN [41]	8.05*	—
CRASM [21]	6.68*	80/48.45%*
HORSD [14]	6.8*	—
LBP-I-RCPR	7.35	80/51.4%
POSE-RCPR	7.64	80/52.4%
RICPR	6.64	80/54.6%
Human [12]	5.6	—

NME and occlusion detection. \* indicates that the result is from the published paper

We also show the CED curve of the COFW data set, as shown in Fig. 4. The comparison of the proposed schemes also proves the superiority of the scheme for occluded face images.

**Occlusion Detection:** Since the COFW dataset provides the basic facts of occlusion, we evaluated the occlusion detection on COFW and compared the scheme with RCPR [12], HPM [37], CoR [27], RPP [18] and CRAM [21] for comparison. The occlusion prediction schemes in Table 2 and Fig. 5 are also superior to existing methods in occlusion detection.

**Fig. 4** CED curves

**Fig. 5** Occlusion results

**Run Time:** We recorded the speeds at 5.3 fps, 4.1 fps and 4.0 fps. We can find that some time correlations are presented. Speed can be increased by implementing it in C++ or using a powerful server. We try to improve the performance of the proposed scheme in the future, for example, by reducing the number of images used for texture correlation, that the proposed some time the correlation.

## 5 Conclusion

In this paper, a robust scheme to solve the sensitive problem for the pose regression approach through jointly analyzing texture and pose of a testing face. By examining the correlation of local binary patterns histograms between the testing face and the training faces, the texture correlated shapes are selected instead of random shapes. At the same time, the pose correlated initialization the robustness of the initialization by estimating the face pose. The scheme obtains remarkably higher accuracies on both facie and occlusion on facial images than the state-of-the-art benchmarks. Moreover, since the initialization is usually independent with facial land- mark localization, the proposed initialization other algorithms.

## References

- Yi, J., Xingyan., G., Yidong, L., Junliang, X., Hui, T.: Towards stabilizing facial landmark detection and tracking via hierarchical filtering: a new method. *J. Franklin Inst.* **357**(5), 3019 (2020)
- Shuo, Y., Ping, L., Chen-Change, L., Xiaoou, T.: From facial parts responses to face detection: a

- deep learning approach. In: IEEE International Conference on Computer Vision, pp. 3676–3684 (2015)
- 3. Renliang, W., Jiwen, L., Yap-Peng, T.: Robust point set matching for partial face recognition. *IEEE Trans. Image Process.* **25**(3), 1163 (2016)
  - 4. Hubin, L., Di, H., Jean-Marie, M., Yunhong, W., Liming, C.: Towards 3D face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *Int. J. Comput. Vision* **113**(2), 128 (2015)
  - 5. Ying, T., Jian, Y., Yigong, Z., Lei, L., Jianjun, Q., Yu, C.: Face recognition with pose variations and misalignment via orthogonal procrustes regression. *IEEE Trans. Image Process.* **25**(6), 2673 (2016)
  - 6. Yongqiang, L., Shangfei, W., Yongping, Z., Qiang, J.: Simultaneous facial feature tracking and facial expression recognition. *IEEE Trans. Image Process.* **22**(7), 2559 (2013)
  - 7. Kamarol, S., Jaward, M., Kälviäinen, H., Parkkinen, J., Parthiban, R.: Pattern Recogn. **92**, 25 (2017).
  - 8. Wei, Z., Youmei, Z., Lin, M., Jingwei, G., Shijie, G.: Multimodal learning for facial expression recognition. *Pattern Recogn.* **48**(10), 3191 (2015)
  - 9. Zhu, X., Ramanan, D.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886 (2012)
  - 10. Ren, S., Cao, X., Wei, Y., Sun, J.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692 (2014)
  - 11. Zhu, S., Li, C., Change Loy, C., Tang, X.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4998–5006 (2015)
  - 12. Burgos-Artizzu, X.P., Perona, P., Dollar, P.: IEEE International Conference on Computer Vision, pp. 1513–1520 (2013)
  - 13. Jourabloo, A., Liu, X.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4188–4196 (2016)
  - 14. Xing, J., Niu, Z., Huang, J., Hu, W., Zhou, X., Yan, S.: *IEEE Trans. Pattern Anal. Mach. Intell.* **99**, 1 (2017)
  - 15. Dollár, P., Welinder, P., Perona, P.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1078–1085 (2010)
  - 16. Cao, X., Wei, Y., Wen, F., Sun, J.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2887–2894 (2012)
  - 17. Kazemi, V., Sullivan, J.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
  - 18. Yang, H., He, X., Jia, X., Patras, I.: *IEEE Trans. Image Process.* **24**(8), 2393 (2015)
  - 19. Tzimiropoulos, G.: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3659–3667 (2015)
  - 20. Liu, Q., Deng, J., Tao, D.: *IEEE Trans. Image Process.* **25**(2), 700 (2016)
  - 21. Liu, Q., Deng, J., Yang, J., Liu, G., Tao, D.: *IEEE Trans. Image Process.* **26**(2), 797 (2017)
  - 22. Lee, D., Park, H., Yoo, C.D.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4204–4212 (2015)
  - 23. Smith, B.M., Dyer, C.R.: CoRR abs/1611.01584. <http://arxiv.org/abs/1611.01584> (2016)
  - 24. Chandran, P., Bradley, D., Gross, M., et al.: Attention-driven cropping for very high resolution facial landmark detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
  - 25. Zhu, M., Shi, D., Zheng, M., et al.: Robust facial landmark detection via occlusion-adaptive deep network. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
  - 26. Zhang, J., Kan, M., Shan, S., Chen, X.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3428–3437 (2016)
  - 27. Yu, X., Lin, Z., Brandt, J., Metaxas, D.N.: European Conference Computer Vision, pp. 105–118 (2014)
  - 28. Seshadri, K., Savvides, M.: *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(10), 2110 (2016)
  - 29. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: *IEEE Signal Process. Lett.* **23**(10), 1499 (2016)
  - 30. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Computer Vision—ECCV (2012)

31. Lazebnik, S., Perona, P., Sato, Y., Schmid, C.: Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 679–692 (2012)
32. Ghiasi, G., Fowlkes, C.C.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1899–1906 (2014)
33. Ojala, T., Pietikainen, M., Harwood, D.: *Pattern Recogn.* **29**(1), 51 (1996)
34. Ojala, T., Pietikainen, M., Maenpaa, T.: *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971 (2002)
35. Pearson, K.: *Proc. Roy. Soc. Lond.* **58**, 240 (1895)
36. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151 (2011)
37. Pan, Y., Zhou, J., Gao, Y., Xiang, J., Xiong, S., Yang, Y.: IEEE International Conference on Automatic Face Gesture Recognition, pp. 619–625 (2017)
38. Lepetit, V., Moreno-Noguer, F., Fua, P.: *Int. J. Comput. Vision* **81**(2), 155 (2009)
39. Zhang, B., Gao, Y., Zhao, S., Liu, J.: *IEEE Trans. Image Process.* **19**(2), 533 (2010)
40. Xiong, X., De la Torre, F.: IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539 (2013)
41. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(5), 918 (2016)

# An Accurate Visual Navigation Method for Wheeled Robot in Unstructured Outdoor Environment Based on Virtual Navigation Line



Zhen Liang , Tiyu Fang , Zihao Dong , and Jinping Li

**Abstract** For the navigation problem in unstructured outdoor environment such as the field without artificial marks, a precise vision navigation method for wheeled robot based on virtual navigation line is proposed. Virtual navigation line is a virtual line based on a small amount of visual information, which can be used in many complex environments. Firstly, the robot needs to design the shape of the walking route. And the general robot's walking routes can be divided into linear route and curve route. Secondly, different virtual navigation line strategies are generated for different walking routes. In the linear navigation part, the virtual navigation line is determined according to the tracking target and the reference target generated. In the curve navigation part, the curve virtual navigation line is fitted according to the multiple targets. Then, the virtual calibration line is determined based on the camera pitch angle, field of view (FOV) and the horizon position in the image, the offset angle and offset distance are computed by using the geometric relationship between virtual calibration line and virtual navigation line. Finally, the fuzzy PID control method is applied to correct robot's direction. The Pioneer3-DX robot is used to do experiments in the outdoor field. The results show that our method can make the robot walk along the designed route in the unstructured outdoor environment accurately, and the navigation accuracy is within 10 cm.

**Keywords** Robot navigation · Visual navigation · Virtual navigation line · Virtual calibration line

---

Z. Liang · T. Fang · Z. Dong · J. Li ()

School of Information Science and Engineering, Jinan 250022, China

e-mail: [ise\\_ljp@ujn.edu.cn](mailto:ise_ljp@ujn.edu.cn)

Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China

Shandong College and University Key Laboratory of Information Processing and Cognitive Computing in 13th Five-Year, Jinan 250022, China

## 1 Introduction

With the rapid development of robot, outdoor navigation has great application prospects and research value, which is one of the key technologies for intelligent mobile robots. The navigation environments of robot can be classified into two categories: structured environments and unstructured environments [1]. For unstructured environments, researchers have proposed many navigation methods. GPS (Global Positioning Systems) and the standalone cellular systems are commonly used by scholars [2], but these methods have low accuracy. RTK (Real-time kinematic) GPS can realize centimeter positioning, but it is susceptible to environment. For example, high buildings, trees and tunnels may block the GPS signal [3]. Therefore, robots often rely on other sensors to achieve accurate navigation.

There are kinds of sensors for robot navigation, [4, 5] present laser-based pose estimation approaches, but the cost of lidar sensors is high. Reference [6] uses ultrasonic sensor for navigation, but this sensor always depends on the temperature, humidity, and so on. References [7–9] employ IMU and odometry data to improve location performance, but its own dead reckoning is subject to drift which may arise due to wheel slip or any measurement errors. Therefore, some scholars fuse a variety of sensors to overcome the defects of a single sensor, such as Refs. [10–15], the system often fuses a stereo-camera sensor, inertial measurement unit, leg odometry, GPS, laser scanner and so on. Although the accuracy of these navigation methods is improved, they often have requirements for the environment and they are usually very expensive.

As a result, many sensors have been considered to find a compromise between accuracy and cost, even in challenging environments. Recently, it has been proven that vision could be a promising navigation sensor. Cameras have the advantage of providing an extensive amount of information while having low weight, limited power consumption, small size and reasonable cost [16, 17]. At present, vision-based navigation methods in unstructured environments can be divided into landmark detection, road detection and vSLAM.

Firstly, Bürki et al. [18] selects some useful landmarks and matches them with previously recorded maps to locate the robot. However, this method is not applicable in the environment without obvious marking. Secondly, unstructured environments may have some road information, so some scholars have detected the navigation path. Li et al. [19] uses dark channel prior and vanishing point to detect road. Li et al. [20] detects road based on intrinsic image and vanishing point. Wang et al. [21] uses a color feature model and Hough transformation to recognize navigation path. Zhang et al. [22] proposes a navigation line detection method based on SUSAN (Smallest Univalue Segment Assimilating Nucleus) corner and sequential clustering algorithm. English et al. [13] extracts and tracks the direction and lateral offset of the dominant parallel texture to track crop rows. However, these methods are generally applicable to the environment with obvious road information or crop rows' structure. Thirdly, SLAM is a research hotspot in the field of robot now. But in complex outdoor environments, there are few applications [23]. Lee et al. [24] has done some research

on dynamic feature extraction, although some problems have also been solved, the dynamic SLAM is still considered as the most difficult problems in the SLAM field. Therefore, vSLAM is faced with problems such as large computation, dynamic changes of environment and so on, the application effect in outdoor unstructured environments is not good.

Therefore, we propose a navigation method for unstructured outdoor environment without artificial marks or road information. Firstly, we construct the virtual navigation line by visual information. Then system determines the offset parameters by the geometric relationship between the virtual navigation line and the virtual calibration line. Finally, the fuzzy PID control method is used to correct the route, so as to achieve the accurate navigation of the robot.

The main innovations of this paper are as follows:

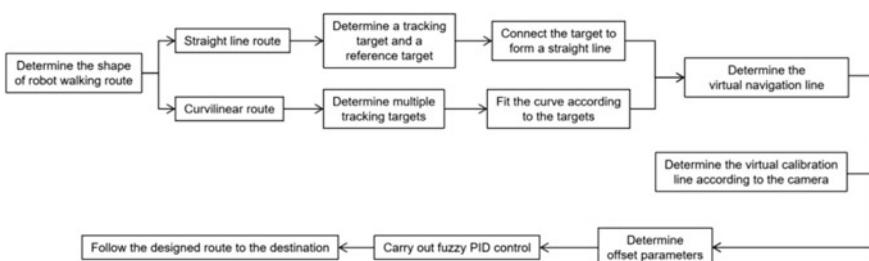
- (1) Our method has no requirements for the environment. It is suitable for many unstructured and complex outdoor environments such as lakes, reed fields, etc.
- (2) The proposed method uses a small amount of visual information to build virtual navigation line, without laying landmarks or real navigation lines.
- (3) The features used in our method are point features rather than line features, so users have more flexibility in route design, which can meet various requirements.

## 2 Principle

### 2.1 General Introduction

The general flow chart of this paper is as follows (Fig. 1).

Our navigation method imitates people's walking rules. We find that when people want to walk in a straight line, they often walk along the lane line or guardrail with the help of road information, or they can find two or more key points in the field of view and walk along the line between the key points. Therefore, in the linear navigation, we take the line between the targets as the virtual navigation line. And when people want to walk in a curve line, they need to determine the key points of the curve in



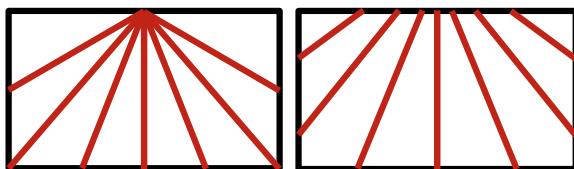
**Fig. 1** Method flow chart

the field of vision, then use the key points to fit a virtual curve, and walk along the curve between the points. So we use a similar method to construct a curved virtual navigation line. Inspired by human walking, we realize the robot's linear and curve walking without artificial landmarks or other obvious road information.

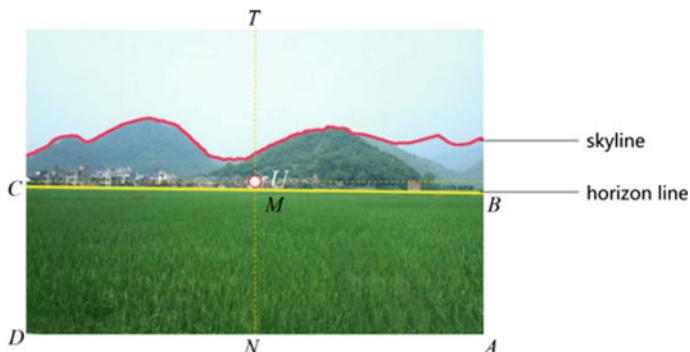
## 2.2 Virtual Calibration Line

When the robot goes straight along the navigation line on the ground, virtual calibration lines are a group of projection straight lines that parallel to the navigation line on the imaging plane [25]. Figure 2 shows two groups of virtual calibration lines converging in different degrees under different pitch angles of camera.

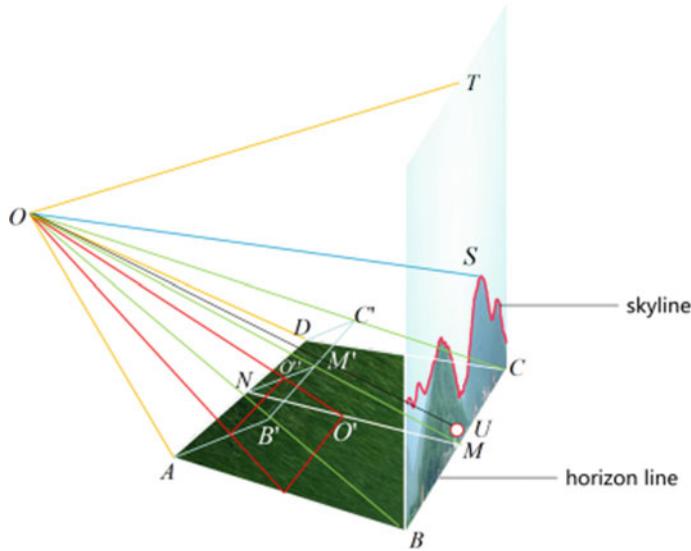
According to the perspective principle, if the pitch angle and the field angle of camera are known, then a group of virtual calibration lines are determined. However, this calculation method can only be applicable when the robot's field of view is all on the ground. When the field of view becomes larger and the view includes the sky and other areas that are not ground, the principle needs to be expanded. Based on the original theory, we add the horizon detection, so the ground area and other areas can be distinguished by horizon, then we can calculate the virtual calibration line by a new mathematical relationship. Figure 3 is a real scene captured by camera, Fig. 4 shows the mathematical geometric relationship in the scene. In Fig. 4, O



**Fig. 2** Navigation lines under different pitch angles



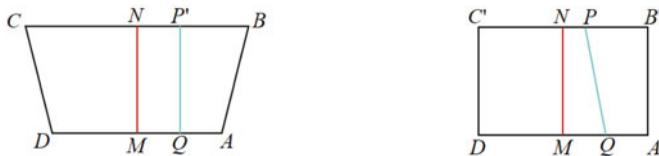
**Fig. 3** The camera's field



**Fig. 4** Schematic diagram of camera imaging

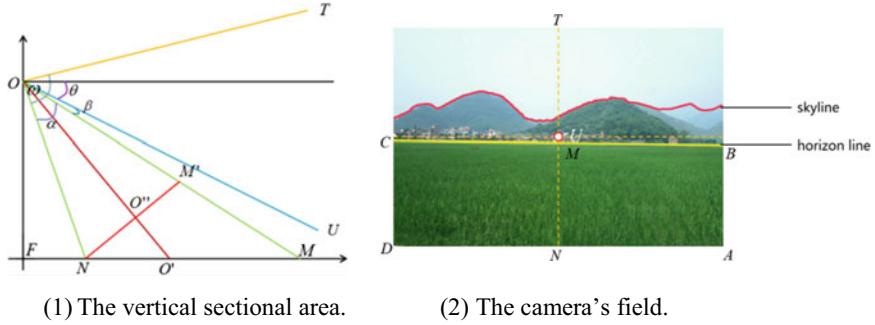
represents the location of camera,  $U$  represents the center of view,  $\angle TOA$  is the field angle. The angle between the optical axis  $OU$  and the horizontal line is the pitch angle. The plane  $ABCD$  is the ground area covered by the camera's field. The ground image obtained by the camera is the projection plane  $AB'C'D'$ , and  $O'$  is the intersection point between the center of the camera's field and the ground, obviously  $AB'C'D \perp OO'$ .

As shown in Fig. 5, the real area of ground view captured by the camera is  $ABCD$ , but the area imaged by the camera is  $AB'C'D'$ . It can be seen that after perspective projection,  $BC$  will be the same width as  $AD$ , and a group of parallel lines on the ground have changed in the imaging plane.  $P'Q$  in Fig. 5a becomes  $PQ$  in Fig. 5b, and  $PQ$  can be considered as a virtual calibration line. Therefore, when  $NP:MQ$  is known, the position of the virtual calibration line can be determined.



(1) Actual capture area of camera. (2) Camera imaging after perspective projection.

**Fig. 5** Comparison of camera imaging and actual capture area



**Fig. 6** Vertical sectional area and camera imaging

From Fig. 4, we can get the vertical sectional area shown in Fig. 6a, O represents the position of camera, OF is the vertical line made downward by the O, intersecting with the extension line of MN at F. As shown in Fig. 5,  $NP:MQ = NP:NP' = B'C':BC = AD:BC$ . As can be seen from Fig. 4,  $AD:BC = AN:BM = B'M':BM = OM':OM$ , so as long as  $OM':OM$  is calculated,  $NP:MQ$  can be obtained.

In Fig. 6a, OU is the optical axis of camera, the pitch angle of the camera is  $\theta$ , the field angle is  $\omega$ , the field angle facing the ground is  $\angle MON$ , we set it to  $\alpha$ , and we set  $\angle UOM$  to  $\beta$ . The pitch angle and the field angle are known from camera,  $\alpha$  and  $\beta$  can be calculated by using the position of the horizon in image. As shown in Fig. 6b,  $\frac{\beta}{\angle TON} = \frac{\beta}{\omega} = \frac{UM}{TN}$ , then  $\beta = \frac{\omega \cdot UM}{TN}$ ,  $\frac{\alpha}{\angle TON} = \frac{\alpha}{\omega} = \frac{MN}{TN}$ , then  $\alpha = \frac{\omega \cdot MN}{TN}$ .

Next, the following derivation can be done.

$$\angle FON = \frac{\pi}{2} - \theta - \beta - \alpha, \quad \angle FOM = \frac{\pi}{2} - \theta - \beta, \quad (1)$$

$$OM' = ON = \frac{OF}{\cos \angle FON} = \frac{OF}{\cos(\frac{\pi}{2} - \theta - \beta - \alpha)}, \quad (2)$$

$$OM = \frac{OF}{\cos \angle FOM} = \frac{OF}{\cos(\frac{\pi}{2} - \theta - \beta)}, \quad (3)$$

$$\frac{OM'}{OM} = \frac{\cos(\frac{\pi}{2} - \theta - \beta)}{\cos(\frac{\pi}{2} - \theta - \beta - \alpha)} = \frac{\sin(\theta + \beta)}{\sin(\theta + \beta + \alpha)}. \quad (4)$$

Therefore, when the pitch angle, field angle and horizon position in view are determined, the virtual calibration lines can be known. In Fig. 5b, starting from any Q, a unique P can be determined, then the corresponding virtual calibration line PQ can be gotten.

## 2.3 Virtual Navigation Line

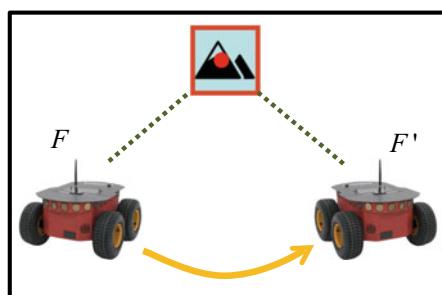
**Virtual Navigation Line in Linear Navigation.** The navigation lines can guide the robot to arrive at destination autonomously, but it is difficult to lay navigation lines in outdoor environment such as fields, so the virtual navigation line is designed to guide the robot to walk. The virtual navigation line is determined by two tracking targets in the field of view, which can assist the robot in determining the direction, and it does not actually exist. The first target is determined according to requirements, and the second target is a reference target automatically generated by the system. The virtual navigation line is constructed based on the target tracking algorithm, after comparative experiments, we select the KCF (Kernel Correlation Filter) [29] as the target tracking algorithm.

In the process of walking, if the robot only walks towards a single target, it can't make sure that the route is straight. As shown in Fig. 7, the robot moves from F to  $F'$ . Although the target is always in the front field of vision, the walking route is curved. Therefore, it is necessary to determine a reference target to reflect the position relationship between the robot and the target, and the relative position of the two targets reflects the change of the robot's walking path.

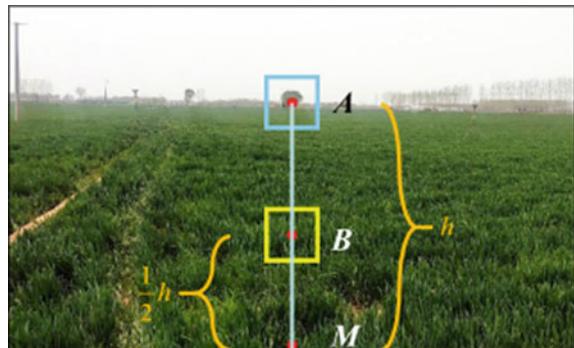
Before determining the lower reference target, we need to control the robot to move until the upper target is move to the center of the field of vision. We assume that the midpoint M (Fig. 5) of the lower boundary of the camera's field of vision represents the position of the robot. We connect the center point A of the initial target that has been moved to the central area with the midpoint M of the lower boundary, the reference target is determined at the midpoint B of the line, and the size of the target is set to be consistent with the upper target, as shown in Fig. 8. If the center of the upper target, the center of the lower target and the position of the robot are in a straight line, then the three points are collinear. If the robot still keeps three points collinear during walking, then the walking path is a straight line.

When the direction of the robot deviates, the center of the upper target, the center of the lower target and the midpoint of the lower boundary of the field of view will no longer be collinear. As shown in Fig. 9, A, B, and M three points are not collinear. When the position of the robot moves from M to  $M'$ , the three points can be collinear again. Therefore, we take the extension line  $AM'$  of the line AB as the virtual

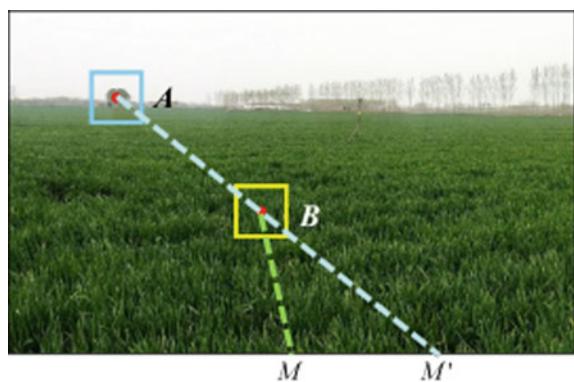
**Fig. 7** Position change of robot with a target



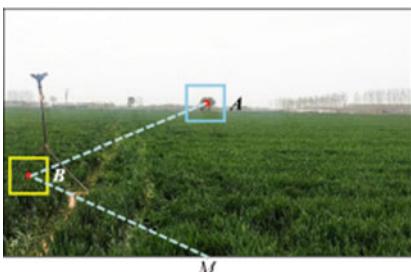
**Fig. 8** Determine the reference tracking target



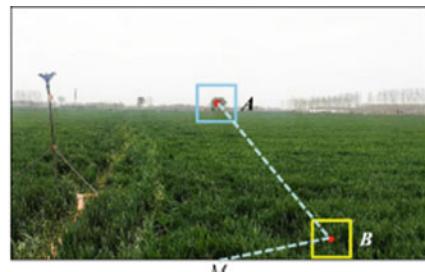
**Fig. 9** The design of virtual navigation line



navigation line. When the deviation occurs, the robot is guided to move in a direction that makes the three points collinear, then the robot can walk straight towards the initial target. It should be noted that in the process of robot walking, the position of targets will change, and the virtual navigation line will change accordingly. When the situation as shown in Fig. 10a and b occurs, the lower target moves out of the



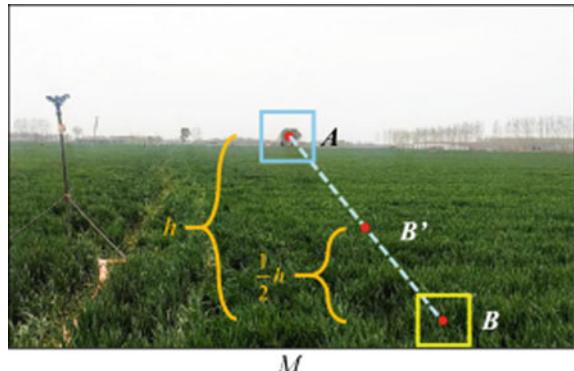
(1) Move out of view horizontally.



(2) Move out of view vertically.

**Fig. 10** Construction of virtual navigation line

**Fig. 11** Update of the lower target



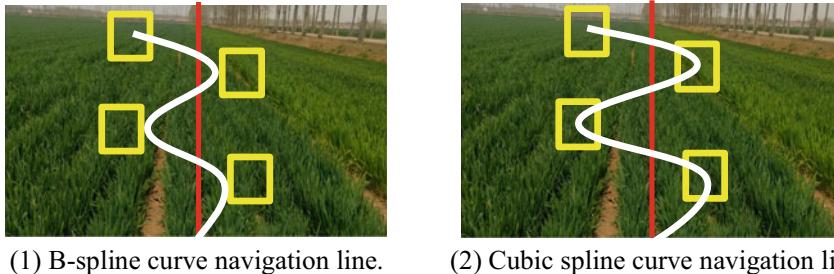
field of view in the vertical or horizontal direction at the next moment, which will not have reference significance, so it is necessary to update the lower reference target, so as to update the virtual navigation line.

The method of updating the virtual navigation line is to record the position of the lower target at this moment when it is about to disappear, as shown in Fig. 11, then connect the lower target (B) with the upper target (A), and set the middle point B' of the line as the center point of the new lower target. And the size of the updated lower target is set to the same as that of the upper target, this update method ensures the stability of the virtual navigation line and the continuity of navigation.

**Virtual Navigation Line in Curve Navigation.** For robots, the walking route is not always linear, there is also the need for curve operation, for example, curve navigation can be used to avoid obstacles. Therefore, we design a curve navigation method for curved operations. The previous work of our laboratory has completed the identification of vertical road signs [26], so when there are road signs, we can find the road signs in the field of view and then fit the curve. We can also set certain rules to find the appropriate key points of the curve in the field of vision. But when there are no obvious signs or road information, we can only select some tracking targets according to the demand. The system will connect them with curves according to the sequence, the first key point is the final destination of the robot, as shown in Fig. 12. The curvilinear virtual navigation line can guide the robot to walk along the curvilinear route. We design two kinds of curve virtual navigation lines, one is B-spline curve approaching but not passing through the points (Fig. 12a), and the other is cubic spline curve passing through the points (Fig. 12b).

**Spline Curve.** B-spline curve is a kind of curve that approximates but not pass through the curve nodes. In the previous section, we have determined N tracking targets, so we have determined the coordinates of N curve nodes. Let  $P_i$  represents the  $i$ th determined point, then the generated B-spline curve can be expressed as Eq. (5).

$$P_{i,n}(t) = \sum_{k=0}^n P_{i+k} \cdot F_{k,n}(t), \quad 0 \leq t \leq 1, i = 0, 1, 2, \dots, m \quad (5)$$



(1) B-spline curve navigation line.

(2) Cubic spline curve navigation line.

**Fig. 12** Curved virtual navigation line

In the formula,  $t$  is the position parameter and  $F_{k,n}(u)$  is called the  $n$  times B-spline basis function. From the expression of B-spline curve, we can know that B-spline curve is defined by segments. If  $m + n + 1$  vertices  $P_i (i = 0, 1, 2, \dots, m + n)$  are given,  $m + 1$  segments  $n$  times parametric curves can be defined, and  $n$  can be any integer from 2 to the number of control points. The specific expression of the basis function is as Eq. (6).

$$F_{k,n}(t) = \frac{1}{n!} \sum_{j=0}^{n-k} (-1)^j \cdot C_{n+1}^j \cdot (t + n - k - j)^n, \quad 0 \leq t \leq 1, k = 0, 1, 2, \dots, n \quad (6)$$

Then the piecewise expression of cubic B-spline curve can be written as Eq. (7).

$$\begin{aligned} P_i(t) &= F_{0,3}(t) \cdot P_i + F_{1,3}(t) \cdot P_{i+1} + F_{2,3}(t) \cdot P_{i+2} + F_{3,3}(t) \cdot P_{i+3}, \\ i &= 0, 1, 2, \dots, m \end{aligned} \quad (7)$$

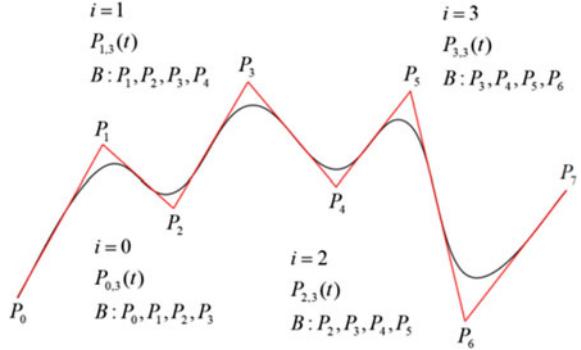
The general matrix form is as follows.

$$P(t) = \sum_{k=0}^3 F_{k,3}(t) \cdot P_k = [t^3 \ t^2 \ t \ 1] \cdot \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} P_0 \\ P_1 \\ P_2 \\ P_3 \end{bmatrix}, \quad 0 \leq t \leq 1 \quad (8)$$

In this formula,  $P_k$  represents the set of vertices of the B feature polygon of the segment curve, and for the  $i$ th curve,  $P_k$  represents four vertices  $P_i, P_{i+1}, P_{i+2}, P_{i+3}$ . A complete cubic B-spline curve defined by  $n$  vertices is connected by  $n-3$  piecewise curves. It can be seen from Fig. 13 that modifying a key point in the cubic B-spline only affects three segment curves, but not all curves, so the shape of the curve has better controllability.

*Cubic spline curve.* Cubic spline curve is a curve fitting method through curve nodes. Because the first and second derivatives are continuous, the curve line is

**Fig. 13** Cubic B-spline curve



relatively smooth, so it has a better shape-preserving function. Cubic spline curve is a curve defined by segments. If data points  $P_i((x_i, y_i)i = 0, \dots, n)$  are given, then on each  $[x_i, x_{i+1}]$ , the expression of cubic spline curve  $S_i(x)$  can be written as Eq. (9).

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, i = 0, 1, \dots, n - 1 \quad (9)$$

In this formula,  $a_i, b_i, c_i, d_i$  represent 4n unknown coefficients. In order to determine the function expression,  $4*n$  conditions should be required. Because cubic spline curve has the characteristics of function continuity, first derivative continuity and second derivative continuity, the following equation can be obtained.

$$\begin{cases} S(x_i) = y_i, i = 0, 1, \dots, n \\ S_-(x_i) = S_+(x_i), i = 1, 2, \dots, n - 1 \\ S'_-(x_i) = S'_+(x_i), i = 1, 2, \dots, n - 1 \\ S''-(x_i) = S''+(x_i), i = 1, 2, \dots, n - 1 \end{cases} \quad (10)$$

According to the value range of i, there are  $4*n-2$  equations, and two more equations are needed to solve the equation. So we add two conditions,  $S''_0(x_0) = 0$ ,  $S''_{n-2}(x_n) = 0$ , which means that the two ends of the curve change gently and are not subject to any force in any direction.

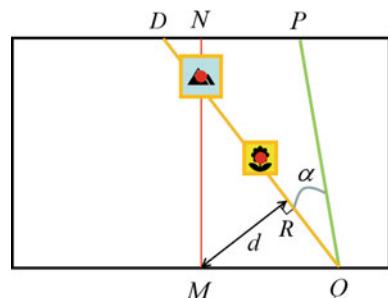
We assume that  $h_i = x_{i+1} - x_i$ ,  $m_i = S''_i(x_i)$ , the equation to be solved can be written as Eq. (11), from which the values of  $a_i, b_i, c_i, d_i$  can be calculated.

$$\begin{bmatrix}
 1 & 0 & 0 & 0 & 0 & \dots & 0 \\
 h_0 & 2(h_0 + h_1) & h_1 & 0 & 0 & \dots & 0 \\
 0 & h_1 & 2(h_1 + h_2) & h_2 & 0 & \dots & 0 \\
 0 & 0 & h_2 & 2(h_2 + h_3) & h_3 & \dots & 0 \\
 \vdots & 0 & 0 & \ddots & \ddots & \ddots & 0 \\
 0 & 0 & 0 & \cdots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\
 0 & \dots & \dots & \cdots & 0 & 0 & 1
 \end{bmatrix} \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \\ \vdots \\ m_n \end{bmatrix} = 6 \begin{bmatrix} 0 \\
 \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\
 \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\
 \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\
 \vdots \\
 \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \\
 0
 \end{bmatrix} \quad (11)$$

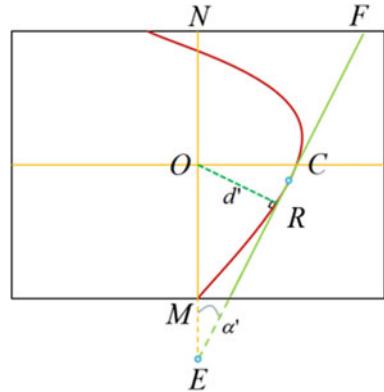
## 2.4 Offset Parameters

**Offset Parameters in Linear Navigation.** According to the geometric relationship between virtual navigation line and virtual calibration line, we can calculate the robot's offset angle and offset distance, and use these offset parameters to make the control decision to ensure the robot walking straight. As shown in Fig. 14, MN is the reference line of the image center, DQ is the navigation line determined by two tracking targets, and PQ is the virtual calibration line determined based on the  $30^\circ$  depression angle. If DQ and PQ coincide, robot will walk along the navigation line. The angle  $\alpha$  formed by DQ and PQ is the offset angle that the robot's moving direction deviates from the navigation line. We set the coordinate of D as  $(x_D, y_D)$ , the coordinate of P as  $(x_P, y_P)$ , and the coordinate of Q as  $(x_Q, y_Q)$ . Then the

**Fig. 14** Determination of offset parameters in linear navigation



**Fig. 15** Determination of offset parameters in curve navigation



calculation of offset angle is shown in the Eq. (12).  $d$  is the distance from the midpoint  $M$  to the navigation line, which represents the offset distance  $MR$  from the robot's dynamic position to the navigation line. If the coordinate of  $M$  is  $(x_M, y_M)$ , the specific calculation of the offset distance is shown in Eq. (13).

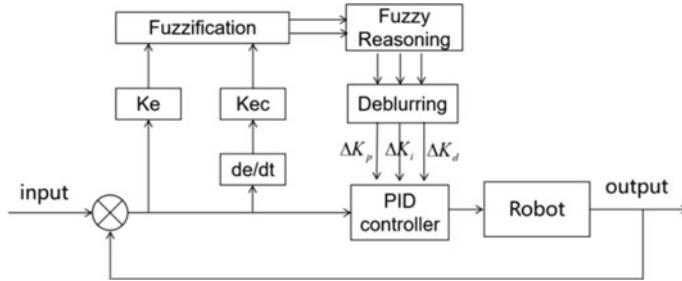
$$\alpha = \arccos \frac{(x_D - x_Q)(x_P - x_Q) + (y_D - y_Q)(y_P - y_Q)}{\sqrt{(x_D - x_Q)^2 + (y_D - y_Q)^2} + \sqrt{(x_P - x_Q)^2 + (y_P - y_Q)^2}} \quad (12)$$

$$d = \frac{|(y_D - y_Q)x_M + (x_Q - x_D)y_M + x_Dy_Q - x_Qy_D|}{\sqrt{(y_D - y_Q)^2 + (x_Q - x_D)^2}} \quad (13)$$

**Offset Parameters in Curve Navigation.** In curve navigation, it is also necessary to obtain the offset distance and offset angle to adjust the robot's motion. As shown in Fig. 15,  $MN$  is a virtual calibration line,  $EF$  is the tangent line of the virtual navigation line at a certain time, which is tangent to the virtual navigation line at  $C$ . The angle between the tangent line and the virtual calibration line is the offset angle  $\alpha'$ . If the coordinate of  $N$  is  $(x_N, y_N)$ ,  $E$  is  $(x_E, y_E)$ , and  $F$  is  $(x_F, y_F)$ , then the calculation of the offset angle is shown in Eq. (14).  $d'$  is the offset distance from the center point  $O$  to the tangent  $EF$ . If the coordinate of  $O$  is  $(x_O, y_O)$ , the calculation of the offset distance is shown in Eq. (15). According to the values of offset angle and offset distance, the fuzzy control table can be established to control the robot.

$$\alpha' = \arccos \frac{(x_N - x_E)(x_F - x_E) + (y_N - y_E)(y_F - y_E)}{\sqrt{(x_N - x_E)^2 + (y_N - y_E)^2} + \sqrt{(x_F - x_E)^2 + (y_F - y_E)^2}} \quad (14)$$

$$d' = \frac{|(y_E - y_F)x_O + (x_F - x_E)y_O + x_Ey_F - x_Fy_E|}{\sqrt{(y_E - y_F)^2 + (x_F - x_E)^2}} \quad (15)$$



**Fig. 16** The flow chart of fuzzy PID

## 2.5 Fuzzy PID Control

Follow Zhang [35], fuzzy PID control method is applied. The adjustment principle of PID control is to calculate the error according to proportion (P), integral (I) and differential (D), and get the output through linear combination to control the object. Fuzzy PID control is based on the PID algorithm, using fuzzy rules for reasoning, querying the fuzzy matrix table for parameter adjustment, to meet the requirements of error (E) and change of error (EC) for parameters self-tuning at different times.

Figure 16 shows the flow chart of fuzzy PID control. We take the offset angle, offset distance and their change rate as the input, and the robot's speed and rotation angle as the output. The general steps are as follows. Firstly, the fuzzy value of the input is obtained by multiplying E and EC by the quantization factor. Secondly, according to the degree of membership function, the membership degree of E and EC are determined. Then, the fuzzy set of output is determined based on the fuzzy rule table. Finally, the output fuzzy value is defuzzified by the defuzzification method, and the final output value is obtained.

## 3 Experiments and Analysis

### 3.1 Hardware Environment

The robot we use is the Pioneer3-DX designed by MobilerRobots Company of the United States. Its controller is H8S series of Hitachi Company and its operating system is ActivMedia Robotics Operating System (AROS). The camera can rotate 360 degrees and pitch 180 degrees up and down on the pan-tilt, as shown in the Fig. 17.

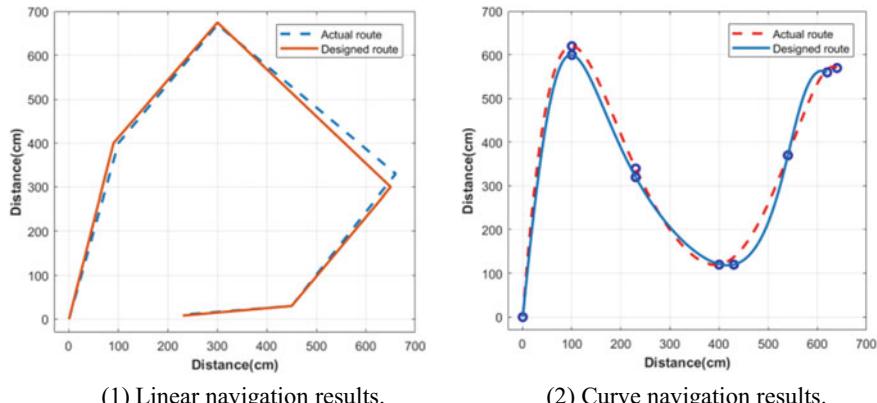


**Fig. 17** Pioneer3-DX

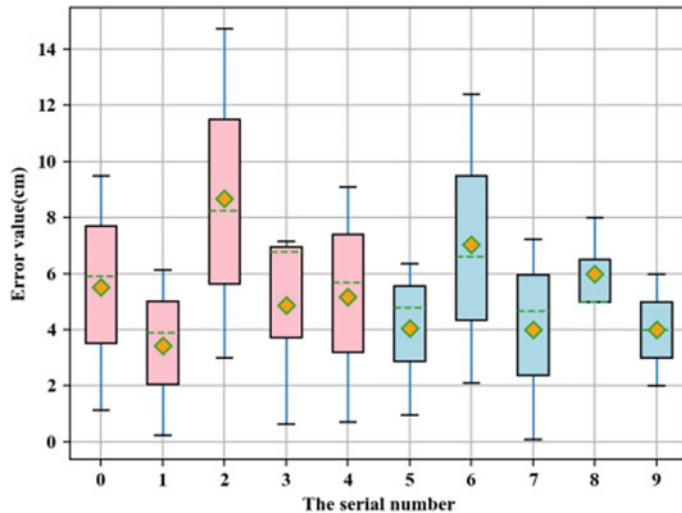
### 3.2 Navigation Effect Analysis

We have done full experiments in the outdoor unstructured environment, and saved the design routes and the actual walking routes based on vertical view. Figure 18 shows two path records of robot, which are straight-line walking path and cubic spline curve walking path.

In the robot walking experiments, we randomly take the results of 5 times of linear walking experiments and 5 times of curve walking experiments, mark some points randomly and record the walking error which is the difference between the designed route and the actual route at the point. The data statistics chart is shown in Fig. 19. We use box chart to do data analysis, show its maximum value, minimum value, lower quartile, upper quartile, average value and median respectively. As can be seen from Fig. 19, the error distance is approximately within 10 cm.



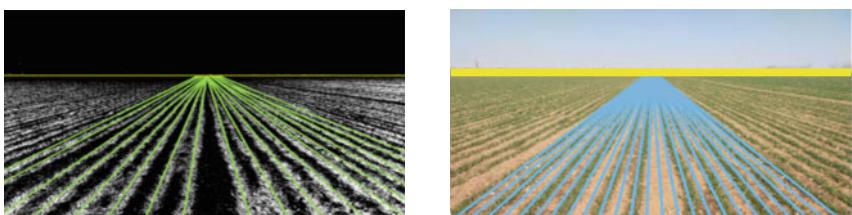
**Fig. 18** Comparison between actual route and design route



**Fig. 19** Statistical chart of error distance

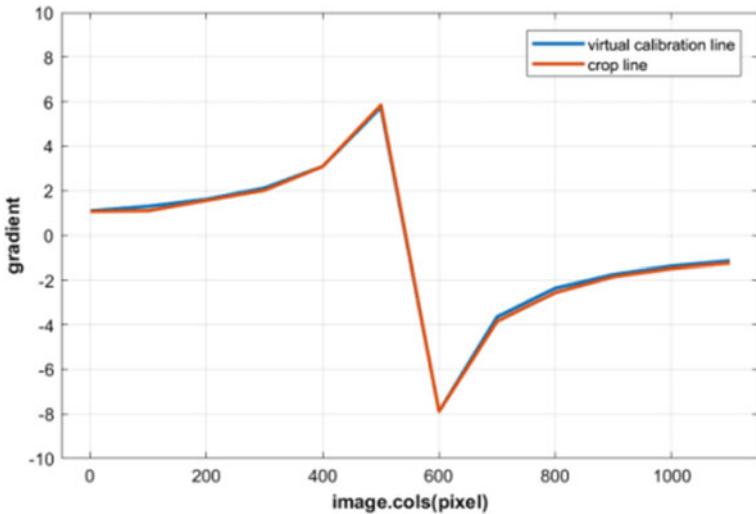
### 3.3 Accuracy Analysis of Virtual Calibration Line

Virtual calibration lines are a group of parallel lines on the ground in the camera view, which is important to guide the robot to walk. We find a farm with clear rows of crops that parallel to each other. Through a series of image operations, we have made clear the straight line of crop lines. Then we manually describe the lines on the ground in the camera's field of vision (Fig. 20a). Then we generate a set of virtual calibration lines (Fig. 20b) by our algorithm and compare them with crop lines. We record the manually marked line, fix the coordinates of the point at the bottom of the image on the line, then calculate the corresponding virtual calibration line from the point, and verify the accuracy of the virtual calibration line by comparing the slopes of the two lines. The statistical results of the slope comparison are shown in Fig. 21,



(1) Crop lines labeling manually. (2) Virtual calibration lines by algorithm.

**Fig. 20** Accuracy analysis of virtual calibration lines



**Fig. 21** The comparison results of the slope

it can be seen that the coincidence degree of virtual calibration lines and crop lines is very high.

### 3.4 Accuracy Analysis of Virtual Navigation Line

In the past, our lab did some work about linear navigation. Zhao [27] laid navigation lines on the ground and Yang [26] made vertical road signs on the wall. We do 20 experiments for each navigation method to compare them. During the robot moving, we record the offset distance and offset angle every 20 cm, and record the average value as this time experimental result. We make a bar chart to show the comparison results. As can be seen from Fig. 22, the effect of laying ground navigation line is the best, followed by our method, and the worst is the vertical landmark method, which shows that our method has a certain feasibility, and there is little difference with the navigation effect of laying ground navigation line.

### 3.5 Accuracy of Target Tracking

In our method, virtual navigation line is formed by target tracking, so the accuracy of target tracking is very important for navigation. Because the environment faced by robot is complex and unknown, some methods based on deep learning are not suitable for robot navigation due to the requirements for samples. We have selected

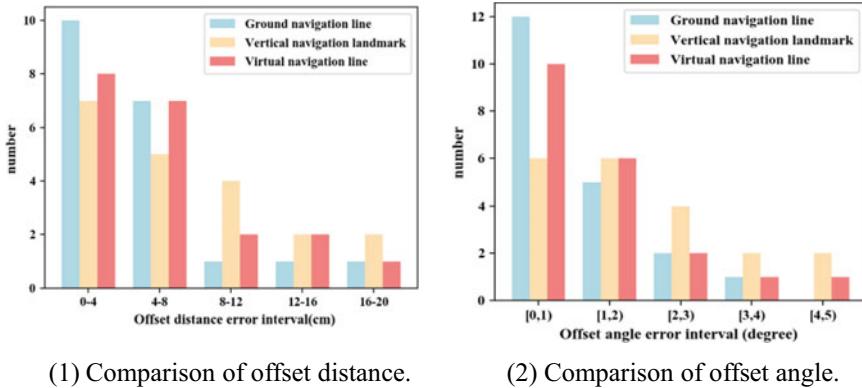


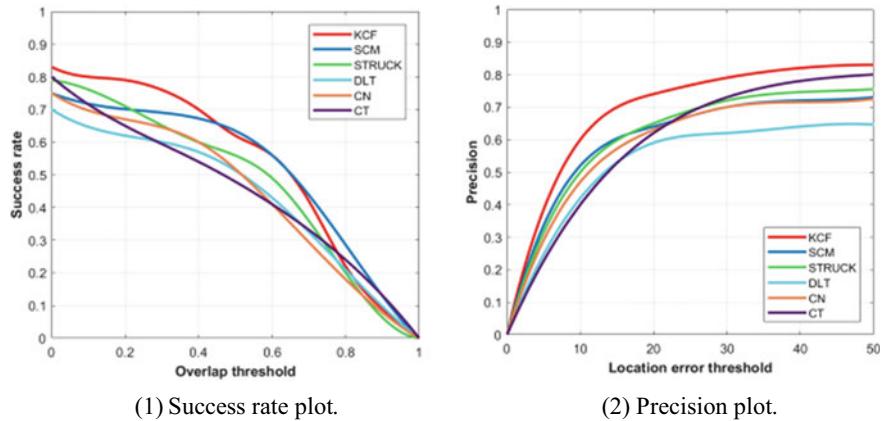
Fig. 22 Comparison of navigation methods

some traditional tracking algorithms for experiment, including KCF, SCM [31], etc. KCF is a discriminant tracking method. It uses the cyclic matrix around the target to collect positive and negative samples, applies ridge regression to train the target detector, and uses the properties of Fourier space to perform operations, so that the algorithm can meet real-time requirements.

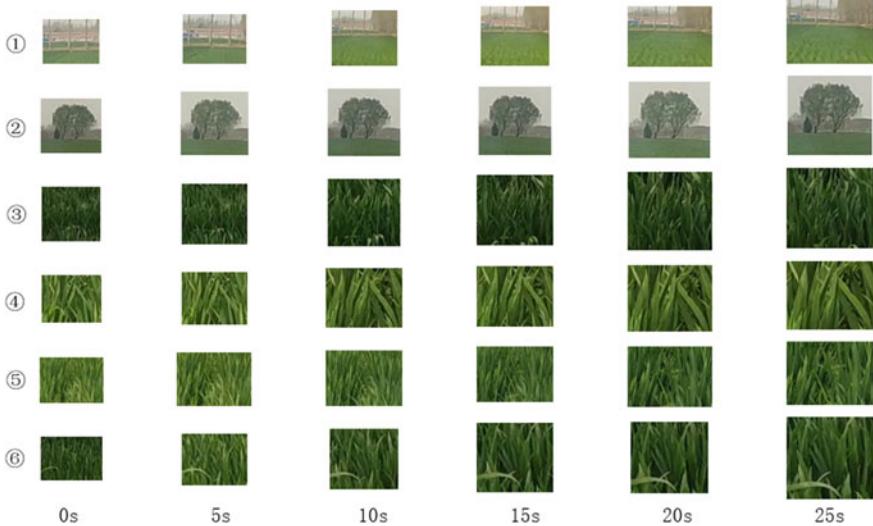
We have done a series of experiments in unstructured environment such as fields and farms to investigate the actual effect of the target tracking algorithm employed in this paper, and analyzed them from quantitative evaluation and qualitative evaluation.

**Quantitative Evaluation.** For quantitative evaluations, we present the results by using success rate (SR) and precision [28]. A tracking result in a frame is considered successful if  $S = \text{Area}(B_T \cap B_G)/\text{Area}(B_T \cup B_G) > \theta$  for a threshold  $\theta \in [0, 1]$ , where  $B_T$  is the tracked bounding box and  $B_G$  denotes the ground truth. SR is defined as the percentage of frames with  $S > \theta$ . The  $\theta$  is set to 0.5 in this paper. The precision plot illustrates the percentage of frames whose tracked locations are within the given threshold distance to the ground truth. Following [28], the threshold value is set at 20 pixels. The evaluated trackers include KCF [29], CT [30], SCM [31], CN [32], STRUCK [33] and DLT [34]. As these results show (Fig. 23), KCF achieves comparable performance in accuracy among all the methods compared.

**Qualitative Evaluation.** We have shown the changes done experiments in some unstructured environments and randomly selected some intermediate results of target tracking to show. As shown in Fig. 24, ①–⑥ of different targets with time, the columns represent the images of different target regions at the same time, and the rows represent the images of the same target region at different times. Figure 25 shows the change of robot's camera images in linear navigation. From these two figures, we can see that as the robot walks, the target size becomes larger and the tracking effect is relatively accurate, so KCF has good tracking performance.



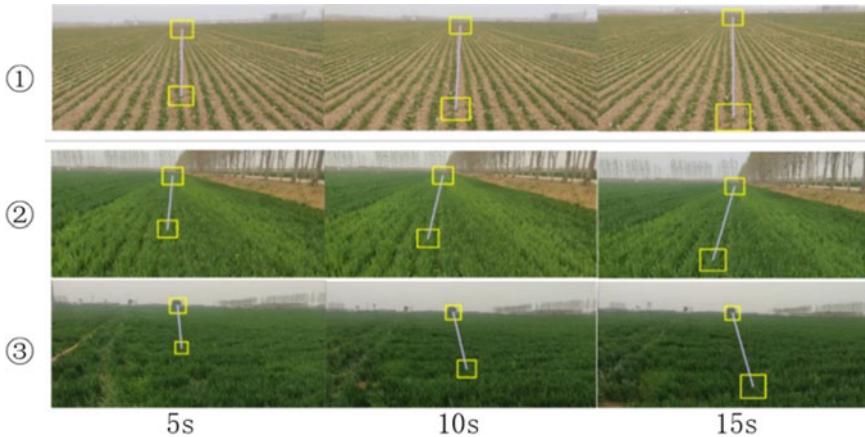
**Fig. 23** Comparison of tracking algorithms



**Fig. 24** The change of targets

## 4 Conclusion

In order to solve the problem of navigation in unstructured outdoor environment such as the field without artificial marks, we propose an accurate visual navigation method for wheeled robot based on virtual navigation line. Firstly, the system needs to determine the virtual calibration line according to the camera's field of view, pitch angle and the position of horizon in image, and then determine the robot's walking route shape, which is divided into linear navigation and curve navigation. Secondly,



**Fig. 25** The changes of robot's view in linear navigation

different virtual navigation lines are generated according to different walking routes. Then, the offset parameters are gotten by using the geometric relationship between virtual calibration line and virtual navigation line. Finally, the fuzzy PID control algorithm is applied to control the robot to walk until it reaches the destination. In this paper, the Pioneer3-DX robot is used to carry out experiments in field. The results show that our method can make the robot walk accurately along design routes in the unstructured outdoor environment, and the navigation accuracy is within 10 cm, which can meet the needs of ordinary civil. However, due to the use of vision sensor, our method is easily affected by the lighting conditions, it needs further improvement in the future.

## References

1. DeSouza, G.N., Kak, A.C.: Vision for mobile robot navigation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(2), 237–267 (2002)
2. Du, H., Zhang, C., Ye, Q., Xu, W., Kibenge, P.L., Yao, K.: A hybrid outdoor localization scheme with high-position accuracy and low-power consumption. *EURASIP J. Wirel. Commun. Netw.* **2018**(1), 1–13 (2018)
3. Sósnička, K., Bury, G., Zajdel, R., Strugarek, D., Drozdzewski, M., Kazmierski, K.: Estimating global geodetic parameters using SLR observations to Galileo, Glonass, Beidou, GPS, and QZSS. *Earth Planets Space* **71**(1), pp. 1–11 (2019)
4. Liu, X., Cao, Z., Jiao, J., Ai, K., Tan, M.: Robot pose estimation and navigation based on the understanding of laser landmarks in unknown environments. In: 11th IEEE International Conference on Control & Automation (ICCA), pp. 332–335. IEEE (2014)
5. Veronese, L.d.P., Aut Cheein, F., Bastos-Filho, T., Ferreira De Souza, A., de Aguiar, E.: A computational geometry approach for localization and tracking in GPS-denied environments. *J. Field Robot.* **33**(7), 946–966 (2016)

6. Varghese, J.Z., Boone, R.G., et al.: Overview of autonomous vehicle sensors and systems. In: International Conference on Operations Excellence and Service Engineering, pp. 178–191 (2015)
7. Alatise, M.B., Hancke, G.P.: Pose estimation of a mobile robot based on fusion of IMU data and vision data using an extended Kalman filter. *Sensors* **17**(10), 2164 (2017)
8. Cheng, Y.H., Meng, Q.H., Liu, Y.J., Zeng, M., Xue, L., Ma, S.G.: Fusing sound and dead reckoning for multi-robot cooperative localization. In: 2016 12th World Congress on Intelligent Control and Automation (WCICA), pp. 1474–1478. IEEE (2016)
9. Wang, S., Deng, Z., Yin, G.: An accurate GPS-IMU/DR data fusion method for driver-less car based on a set of predictive models and grid constraints. *Sensors* **16**(3), 280 (2016)
10. Ma, J., Bajracharya, M., Susca, S., Matthies, L., Malchano, M.: Real-time pose estimation of a dynamic quadruped in GPS-denied environments for 24-hour operation. *Int. J. Robot. Res.* **35**(6), 631–653 (2016)
11. Biber, P., Weiss, U., Dorna, M., Albert, A.: Navigation system of the autonomous agricultural robot bonirob. In: Workshop on Agricultural Robotics: Enabling Safe, Efficient, and Affordable Robots for Food Production (Collocated with IROS 2012), Vilamoura, Portugal (2012)
12. English, A., Ball, D., Ross, P., Upcroft, B., Wyeth, G., Corke, P.: Low cost localisation for agricultural robotics. In: Proceedings of the 2013 Australasian Conference on Robotics and Automation, pp. 1–8. Australasian Robotics and Automation Association (ARAA) (2013)
13. English, A., Ross, P., Ball, D., Corke, P.: Vision based guidance for robot navigation in agriculture. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 1693–1698. IEEE (2014)
14. English, A., Ross, P., Ball, D., Upcroft, B., Corke, P.: Learning crop models for vision-based guidance of agricultural robots. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1158–1163. IEEE (2015)
15. Bergerman, M., Maeta, S.M., Zhang, J., Freitas, G.M., Hamner, B., Singh, S., Kantor, G.: Robot farmers: autonomous orchard vehicles help tree fruit production. *IEEE Robot. Autom. Mag.* **22**(1), 54–63 (2015)
16. Ben-Afia, A., Deambrogio, L., Sal'os, D., Escher, A.C., Macabiau, C., Soulier, L., Gay-Bellile, V.: Review and classification of vision-based localisation techniques in unknown environments. *IET Radar Sonar Navig.* **8**(9), 1059–1072 (2014)
17. Winterhalter, W., Fleckenstein, F., Dornhege, C., Burgard, W.: Localization for precision navigation in agricultural fields beyond crop row following. *J. Field Robot.* **38**(3), 429–451 (2021)
18. Bürki, M., Cadena, C., Gilitschenski, I., Siegwart, R., Nieto, J.: Appearance-based landmark selection for visual localization. *J. Field Robot.* **36**(6), 1041–1073 (2019)
19. Li, Y., Ding, W., Zhang, X., Ju, Z.: Road detection algorithm for autonomous navigation systems based on dark channel prior and vanishing point in complex road scenes. *Robot. Auton. Syst.* **85**, 1–11 (2016)
20. Li, Y., Tong, G., Sun, A., Ding, W.: Road extraction algorithm based on intrinsic image and vanishing point for unstructured road image. *Robot. Auton. Syst.* **109**, 86–96 (2018)
21. Wang, P., Meng, Z., Luo, C., Mei, H.: Path recognition for agricultural robot vision navigation under weed environment. In: International Conference on Computer and Computing Technologies in Agriculture, pp. 242–248. Springer (2013)
22. Zhang, Q., Chen, M.S., Li, B.: A visual navigation algorithm for paddy field weeding robot based on image understanding. *Comput. Electron. Agric.* **143**, 66–78 (2017)
23. Gao, X., Li, J., Fan, L., Zhou, Q., Yin, K., Wang, J., Song, C., Huang, L., Wang, Z.: Review of wheeled mobile robots navigation problems and application prospects in agriculture. *IEEE Access* **6**, 49248–49268 (2018)
24. Lee, C.S., Clark, D.E., Salvi, J.: Slam with dynamic targets via single-cluster PHD filtering. *IEEE J. Sel. Top. Sign. Process.* **7**(3), 543–552 (2013)
25. Li, J., Xiu, Z., Lv, Y., Han, Y.: A patrol robot visual navigation method based on virtual calibration line. Ph.D. thesis (2011)

26. Yang, S., Cai, F., Zhao, P., Han, Y., Li, J.: A visual self-localization method of patrol robot based on vertical highlighted landmarks. *J. Nanjing Normal Univ (Eng)* **3**, (2019)
27. Zhao, P.: Vision-based navigation and voice information services by using patrol robot. Master's thesis, University of Jinan (2018)
28. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2411–2418 (2013)
29. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 583–596 (2014)
30. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: European Conference on Computer Vision (ECCV), pp. 864–877 (2012)
31. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1838–1845. IEEE (2012)
32. Danelljan, M., Shahbaz Khan, F., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1090–1097 (2014)
33. Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.M., Hicks, S.L., Torr, P.H.: Struck: structured output tracking with kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(10), 2096–2109 (2015)
34. Wang, N., Yeung, D.Y.: Learning a deep compact image representation for visual tracking. In: Advances in Neural Information Processing Systems (2013)
35. Zhang, C.: Design and fuzzy PID control of hexapod walking robot system based on STM32. Master's thesis, Zhejiang Sci-Tech University (2016)

# DABU-Net: Dilated Convolution and Attention U-Net with Boundary Augment for Medical Image Segmentation



Ye Yuan , Yajing An , and Guoqiang Zhong

**Abstract** U-Net has a good representation learning capability in medical image segmentation because it can extract contextual information from an image. However, U-Net also has two shortcomings. First, U-Net can only extract features on fixed scales, which limits its ability. Second, the feature maps from the same-scale encoder and decoder are semantically dissimilar, so that the shortcut connections of U-Net may cause semantic gap between the low- and high-level layers. In this paper, we propose a deep end-to-end network dubbed Dilated Convolution and Attention U-Net with Boundary Augment for Medical Image Segmentation (DABU-Net). In the encoding path, to increase the receptive field and capture the multi-scale information, we use dilated convolution to design the dilated convolution block (DCblock). In the decoding path, we design the spatial and channel attention block to narrow the semantic gap, and we use sobel operator to enhance the segmentation area. We evaluate the proposed network on three medical image datasets: TCGA Brain MRI dataset, LiTS 2017 liver segmentation dataset, and ISIC 2018 skin lesion segmentation dataset. Experimental results show that the DABU-Net has achieved better performance compared with other methods.

**Keywords** Medical image segmentation · Deep learning · U-Net · Attention

## 1 Introduction

Medical image segmentation is a vital task in medical diagnosing. It can help physicians determine the lesion area, assess the effect before and after treatment. Doctors need long-term professional training, and the results are often affected by doctors' experience, fatigue, and patience. In contrast to natural images, the segmentation of medical images is more difficult, because it requires high accuracy and high stability, and the medical images often have a low signal–noise ratio.

---

Y. Yuan · Y. An · G. Zhong ()

Department of Computer Science and Technology, Ocean University of China, Qingdao, China  
e-mail: [gqzhong@ouc.edu.cn](mailto:gqzhong@ouc.edu.cn)

Deep learning method has made revolutionary progress in many tasks because of its powerful feature representation ability. It has shown strong ability in image processing tasks, and has become an important part of image segmentation. Fully convolutional network (FCN) [1] has achieved a good performance on natural images as a representative work of segmentation. Ronneberger et al. [2] proposed U-Net in 2015, which first applied the idea of skip connection to the segmentation, and obtained the most accurate results at the time. Many variants of U-Net have been proposed recently [3–5]. Zhou et al. [6] think it is not appropriate that skip connection combines the shallow features with the deep features directly, because it will produce a semantic gap. They proposed U-Net++ to improve the skip-connection by adopting dense blocks and deep supervision. In some variants, some processing steps (e.g., attention gates [7]) are used to process the encoder’s feature maps. Azad et al. proposed BCDU-Net [8], by introducing Bi-ConvLSTM into the skip connection and processing feature maps with dense connections, which achieved better performance. Although a lot of work has been put forward, the accuracy of lesion segmentation in medical images still needs improvement.

In this manuscript, we design a new deep learning network named DABU-Net for medical image segmentation. The U-Net uses convolution and pooling layers to improve the receptive field and extract features, which miss the long-distance dependencies. To reduce the loss of accuracy, we propose the DCblock, which introduces dilated convolution to increase the receptive field. In addition, the features extracted by encoders have higher resolution, but the features of same-scale decoders have more semantic information. The skip-connections just concatenate them simply, which may cause the semantic gap between the upper and the lower layer of the U-Net. We use spatial and channel attention to solve this problem. Finally, to make the boundary of segmentation area more accurate, we designed the boundary augment module. We validate the DABU-Net on three different datasets, and also the experiments demonstrate that our model achieves higher performance than alternative ways.

## 2 Related Work

In recent years, methods based on deep learning have significantly increased the performance of image segmentation in natural scenes, and they also have been dominating medical image segmentation tasks. The traditional segmentation method uses pixel blocks as the input of convolutional neural network for training and prediction. However, this requires a lot of computation, because the pixel blocks used are basically overlapped. Meanwhile, the size of the receptive field is restricted by the scale of the pixel blocks.

In order to solve these problems, fully convolutional network (FCN) was proposed by Jonathan et al. [1] in 2015. FCN removes the fully connected layers and replaces it with convolutional layers, and reconstruct the image with deconvolution. At the same time, high-level information and low-level information are combined by using the

shortcut connection to produce accurate segmentation results. However, the results generated by FCN are vague and the details are not very good. The classification of pixels does not consider the relationship between each other.

Larger patches require additional max-pooling layers, which will reduce the accuracy of segmentation results, while tiny patches will reduce the models' ability to capture context information [9]. To solve the trade-off problem between location information and context information, Ronneberger et al. proposed an improved convolutional network named U-Net [2]. U-Net contains an encoder, a bottleneck module, and a decoder. It combines low-resolution information with high-resolution information, making it ideal for medical image segmentation. However, because of the differentiation of organ structure and the diversification of lesion shapes, only using U-Net structure to segment lesions cannot meet the requirements of accuracy and speed. With the development of the residual structure [10], dense module [11], inception module [12], and attention mechanism [13], recent work has added different modules on the U-Net to achieve reasonable segmentation results [3, 5, 14, 15]. However, the accuracy of these methods is not high enough, while a small mistake in the medical field will have a big impact.

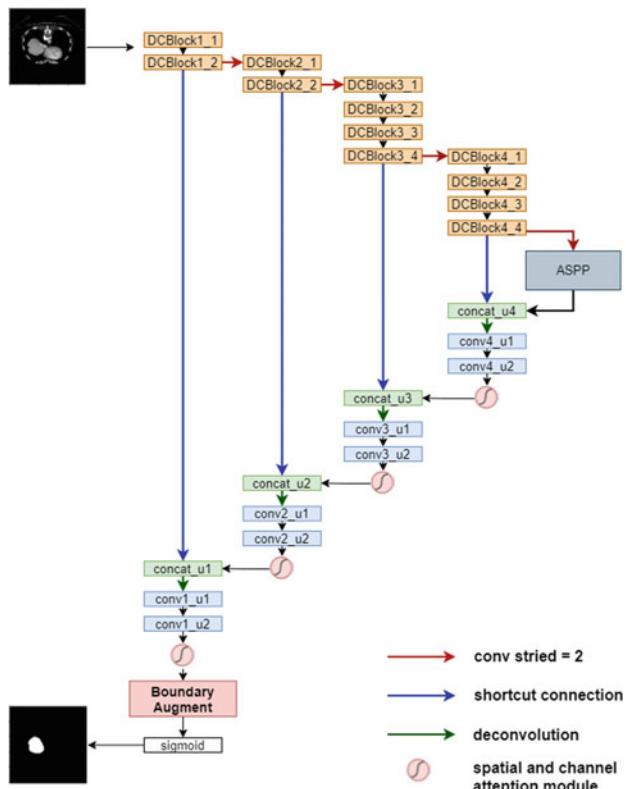
In medical imaging, because the location of the lesion is relatively large and there are many other irrelevant features, it is extremely important to focus on target features and suppress irrelevant features. Squeeze-and-Excitation (SE) [16] is proposed in 2018, which can be stimulated from both space and channel to achieve the effect of enhancing features, Roy et al. [17] introduced three SE structures on the U-Net, which can automatically acquire the importance of each feature channel. Oktay et al. [7] proposed Attention U-net in 2018, and they added an integrated attention gate to adjust the features. An enhanced attention module (AAM) [18] was proposed by Ni et al. to emphasize the target channel by modeling semantic dependencies and extract high-level and low-level context information and semantic features, thus fusing multi-level features and capturing contextual information.

R2U-Net [19] uses residual connection and recurrent convolution instead of the original convolutional layers. This method ensures the depth of the network while reducing the effect of gradient vanishing, which has a significant effect on the extraction of low-level features. However, R2U-Net uses a lot of recurrent convolution, which makes it difficult to train and requires a lot of memory. CE-Net [5] uses residual structure to extract features, and uses DACblocks and RMPblocks to extract multi-scale information. BCDU-Net [8] use Bi-ConvLSTM to extract information from encoder and reuse feature maps with dense module, which achieved good performance. DoubleU-Net [4] concatenate two U-Net structures to improve the performance on some medical image segmentation tasks, which includes two encoders and two decoders.

### 3 Proposed Method

Figure 1 shows an overview of the proposed architecture. The encoding path of DABU-Net consists of four stages, and low-level stage includes two DCblocks and high-level stage includes four DC blocks. We use the dilated convolution to construct the encoder and applied ReLU function as activation function to introduce non-linearity. To get multi-scale information and further increase the receptive field, we replace the last block of encoding path with atrous spatial pyramid pooling (ASPP) [20]. The ASPP samples a given input in parallel with five atrous convolutions at a different sampling rate, which means it helps capture the contextual information at multiple scales and extract high-resolution feature maps.

In decoding path, each block performs transposed convolution as up-sampling. Then these features are concatenated with the shortcut connection features from the encoder. In this way, the spatial dimension of the input feature is doubled, and the



**Fig. 1** Overall architecture of DABU-Net. The network consists of U-Net, DCblock (orange box), spatial and channel attention module, and boundary augment module (pink box)

channel dimension is halved. Finally, these features are fed into two convolutional layers to restore resolution information.

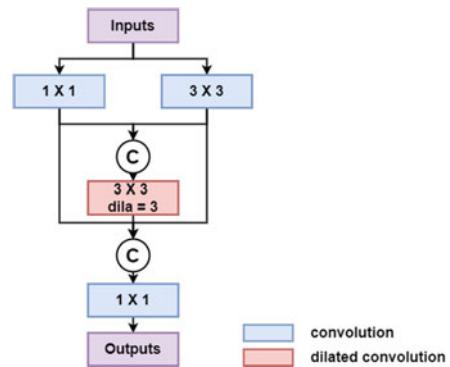
### 3.1 DCblock

Compared with traditional convolution, the dilated convolution has some advantages [21]. It can increase the receptive field of the model without increasing the parameters, and obtain richer contextual information. The feature map generated by dilated convolution can be the same scale as the input, meanwhile, each output neuron has a larger receptive field, so it can encode higher-level semantics.

The architecture of DCblock is illustrated in Fig. 2. In the DCblock, the features are fed into  $1 \times 1$  and  $3 \times 3$  convolutional filters separately, then these features are fed into  $3 \times 3$  dilated convolutional filters. In this way, the DCblock can get different sizes of receptive fields. Finally, all the features are fed into  $1 \times 1$  convolutional filters to reduce the channel dimension. This strategy can make our network extract features more effectively.

Meanwhile, the original U-Net uses max-pooling layers to reduce the scale of the image and increase the receptive field, resulting in the reduction of resolution, and some information will be lost. At this time, when the up-sampling is restored to the original image scale, the segmentation accuracy will be affected. To avoid this problem, convolutional layers that stride equals 2 are used to replace the max-pooling layers, which reduces the image size without losing information.

**Fig. 2** The architecture of our proposed DCblock

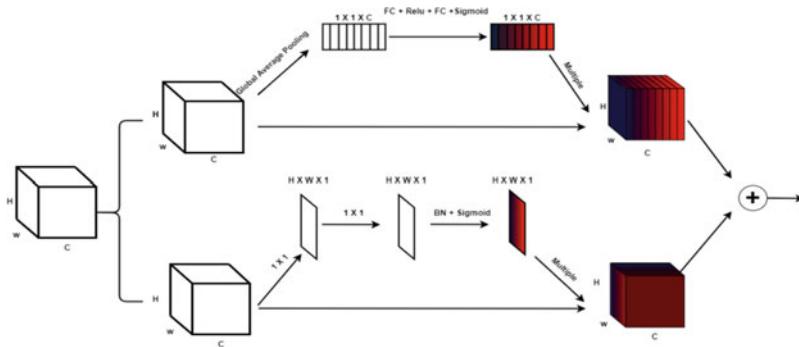


### 3.2 Attention Module

The shortcut connection of U-Net is mainly to fuse the features to better restore resolution information. However, this simple fusion of high- and low-level semantic information easily leads to semantic gap [9]. Therefore, in our model, spatial and channel attention mechanism is used to extract the more important features and narrow the semantic gap caused by a direct concatenation. After concatenated the features from down-sampling layer and up-sampling operation, we send these features to attention module, which structure is shown in Fig. 3. Spatial attention shows the degree of attention differently to different positions but ignores the information of channel domain, while channel attention does the opposite. Thus, we combine these two attention mechanisms to take advantage of their respective strengths for better results.

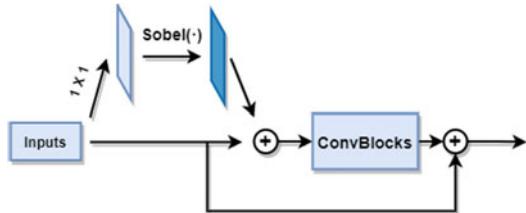
We use the SE block [16] as channel attention, which uses global average pooling as a squeeze step. Then in the excitation step, we use two fully connected layers (FC) to process the results of the squeeze, the first FC compresses the  $C$  channels into  $C/r$  channels to reduce the computation ( $r$  is the ratio of compression), and the second FC is restored back to  $C$  channel. Sigmoid function is used to limit the features to the range of  $[0, 1]$ . This structure controls the weights of the channels to enhance the important features and weak the unimportant features.

Spatial attention mechanism can acquire a spatial feature map to enhance or suppress features at different locations. For spatial attention, the features of different channels are compressed into a same feature map by  $1 \times 1$  convolution, then this feature map is fed into a  $3 \times 3$  convolution to adjust the weights. Batch normalization and sigmoid function are used to limit the features. Finally, this feature map is multiplied by the features of each channel. In every scale, we add these two attention mechanisms to merged features.



**Fig. 3** The architecture of attention module. The above shows the channel attention, the following shows the spatial attention used in our model, and finally the features are added to merge the information

**Fig. 4** The architecture of boundary augment module



### 3.3 Boundary Augment Module

To improve the segmentation ability of lesion area, we design the boundary augment module on our network. As shown in Fig. 1, the features of decoder are fed into boundary augment module to generate more accurate results. The architecture of this module is illustrated in Fig. 4. The features are first fed to  $1 \times 1$  convolution to squeeze features, and then the Sobel operator is used to extract the edge of extracted features. Sobel operator uses a  $3 \times 3$  filter to obtain a gradient image in the horizontal and vertical directions, and is robust to noise. Finally, the boundary information is added to the original features. By undertaking this boundary augment, our network can establish the relationship between region and boundary, thus improving the segmentation accuracy.

## 4 Experiments

In this section, we first describe some details of our experiment. Then, to assess the performance of our proposed DABU-Net, we test it on the TCGA Brain MRI, LiTS 2017, and ISIC 2018 datasets, and compare our DABU-Net with U-Net, Attention U-Net, R2U-Net, CE-Net, BCDU-Net ( $D = 3$ ), and MultiResU-Net, they are published methods based on U-Net for medical image segmentation in the past three years, except U-Net. Finally, we do some ablation experiments to prove the effectiveness of the DCblock, attention module, and boundary augment module.

### 4.1 Configuration and Evaluation Metrics

During the training stage, we employ Adam optimizer with the learning rate of 1e-4 to train the networks, and batch size is set to 8. In addition, we reduce learning rate when the accuracy is not reduced for 10 epochs. The binary cross-entropy loss function was used for all networks. The framework used is Keras, which are implemented on the NVIDIA GeForce GTX 1080Ti platforms.

## 4.2 Results on Brain MRI Dataset

To assess the performance of DBAU-Net, we conducted experiments on Brain MRI images [22]. This dataset corresponds to 110 patients included in The Cancer Genome Atlas lower-grade glioma. The corresponding ground truths are annotated manually by a researcher by drawing an outline on each slice. We totally obtained 2,440 images from these images, and 70, 10, and 20% of these images are used for training data, validation data, and test data, respectively.

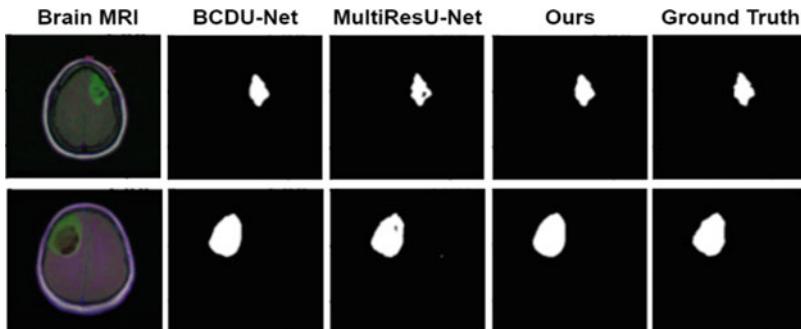
Table 1 shows the quantitative results of Brain MRI dataset. Compared with those methods, our network achieved a large improvement. From Table 1, it can be seen that the F1-score obtained by DABU-Net is 90.05%, which is 2.03% higher than the MultiResU-Net. The promising results indicate the effectiveness of our method. For the result of BCDU-Net, notice that it takes many times as long as we do for the same experiment.

Some segmentation images of brain lesion area are shown in Fig. 5. The first column in figure is the input images of the brain, the second column is the ground truth masks, and the rest columns are segmentation results of our model and other

**Table 1** Results of the proposed network and other advanced methods on TCGA Brain MRI dataset

Methods	Year	F1-score	Sensitivity	Specificity	Accuracy	JS	AUC
U-Net [2]	2015	0.8574	0.8303	0.9965	0.9913	0.7504	0.9134
Attention U-Net [7]	2018	0.8445	0.8317	0.9955	0.9904	0.7308	0.9136
CE-Net [5]	2019	0.8274	0.8079	0.9953	0.9894	0.7056	0.9016
BCDU-Net (D = 3) [8]	2019	0.8774	<b>0.9965</b>	0.8655	0.9924	0.7815	0.9310
MultiResU-Net [3]	2020	0.8847	0.8631	<b>0.9971</b>	0.9929	0.7933	0.9301
DABU-Net (Ours)	–	<b>0.9005</b>	0.9004	0.9968	<b>0.9937</b>	<b>0.8191</b>	<b>0.9486</b>

The best results are highlighted with bold face



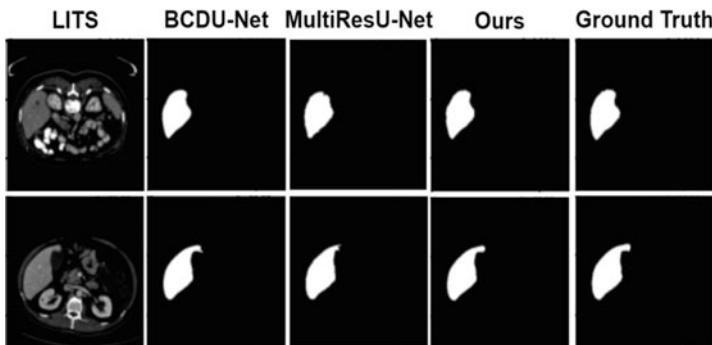
**Fig. 5** Segmentation results of MultiResU-Net, BCDU-Net and our DABU-Net on Brain MRI dataset

competitive models. As can be seen, the segmentation boundary of our models is better than other models.

### 4.3 Results on LiTS 2017

For further comparison, we use the public Liver Tumor Segmentation (LiTS) 2017 dataset to evaluate the proposed DABU-Net. This dataset contains 131 CT scans, each of which has the same size in-plane resolution but a different number of axial slices. The LiTS 2017 dataset contains the ground truth of the liver and the tumor, but we just use the liver data to evaluate the proposed DABU-Net. To do that, we first extract 10 2D-slices from each 3D scan, and the corresponding annotations are processed in the same way, thus we obtained 1310 2D slices of the data.

Figure 6 shows some segmentation outputs of the DABU-Net for liver segmentation, which indicates that our proposed network could successfully extract the liver from an image. As shown in Table 2, our method reached up to F1-score of 0.941 and AUC of 0.954, and performed better than other methods.



**Fig. 6** Segmentation results of BCDU-Net, MultiResU-Net, and our DABU-Net on LiTS 2017 dataset

**Table 2** Results of the proposed network and other advanced methods on LITS 2017 dataset

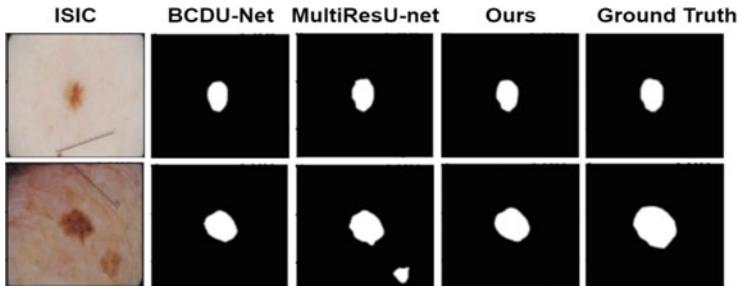
Methods	Year	F1-score	Sensitivity	Specificity	Accuracy	JS	AUC
U-Net [2]	2015	0.9152	0.9010	0.9938	0.9860	0.8437	0.9474
Attention U-Net [7]	2018	0.9181	0.9142	0.9929	0.9864	0.8487	0.9536
R2U-Net [19]	2018	0.9125	0.8870	0.9948	0.9858	0.8390	0.9409
CE-Net [5]	2019	0.9131	0.9062	0.9928	0.9856	0.8401	0.9495
BCDU-Net (D = 3) [8]	2019	0.9279	<b>0.9184</b>	0.9968	0.9902	0.8655	0.9445
MultiResU-Net [3]	2020	0.9109	0.8652	0.9968	0.9859	0.8364	0.9310
DABU-Net (Ours)	–	<b>0.9411</b>	0.9103	<b>0.9977</b>	<b>0.9905</b>	<b>0.8887</b>	<b>0.9540</b>

The best results are highlighted with bold face

#### 4.4 Results on ISIC 2018

The ISIC 2018 is a dataset for skin lesion segmentation. Its training data includes 2,594 images of skin lesions and corresponding annotations. We use 70% of the data as training images, 10% of the data as validation images, and the remaining data as test set. Each image is resized to  $256 \times 256$  in our experiments.

Figure 7 shows the results of these networks on ISIC 2018 dataset, and the quantitative results are shown in Table 3. Compared to those methods, our network achieved a large improvement, especially in F1-Score and Accuracy. From Table 3, it can be seen that DBAU-Net achieves 89.6% in F1-score and 81.3% in JS, which outperforms the BCDU-Net by 4.5% in terms of F1-score and 1.3% in JS. The best result achieved by MultiResU-Net was  $JS = 77.7\%$ . Compared with this result, our network made a great progress.



**Fig. 7** Segmentation results of BCDU-Net, MultiResU-Net, and our DABU-Net on ISIC 2018 dataset

**Table 3** Results of the proposed network and other advanced methods on ISIC 2018 dataset

Methods	Year	F1-score	Sensitivity	Specificity	Accuracy	JS
U-Net [2]	2015	0.647	0.708	0.964	0.890	0.594
Attention U-Net [7]	2018	0.665	0.717	0.967	0.897	0.566
R2U-Net [19]	2018	0.679	0.792	0.928	0.880	0.581
CE-Net [5]	2019	0.852	0.786	0.930	0.935	0.743
BCDU-Net (D = 3) [8]	2019	0.851	0.785	0.982	0.937	0.683
MultiResU-Net [3]	2020	0.874	0.812	<b>0.986</b>	0.945	0.777
DABU-Net(Ours)	–	<b>0.896</b>	<b>0.857</b>	0.982	<b>0.953</b>	<b>0.812</b>

The best results are highlighted with bold face

**Table 4** Segmentation results by ablation study of our methods on the ISIC 2017 dataset

Model	F1-score	Jaccard score
U-Net	0.647	0.594
U-Net + ASPP	0.872	0.773
U-Net + DC block	0.887	0.798
U-Net + attention module	0.873	0.775
U-Net + boundary augment	0.861	0.757
All	0.896	0.812

#### 4.5 Ablation Study

To further evaluate the effectiveness of the proposed DCblock, boundary augment module, and attention module, we conducted the ablation studies using the ISIC 2018 dataset as example. We use U-Net as the baseline model, and add these modules to compare with baseline model.

In the original U-Net, low-level features contain higher resolution information, such as texture and color, while high-level features contain more structural semantic information. This simply concatenate between low-level features and high-level features expressed by images will lead to a semantic gap. In our DABU-Net, we use attention module to combine these features. In addition, we employ the dilated convolution to enhance the encoder blocks, aiming at enhancing the learning capability. We then tested the baseline, U-Net with DC block, attention module, boundary augment module, and ASPP respectively. The ablation results are shown in Table 4. It shows a better segmentation result of the proposed module than the original U-Net.

### 5 Conclusion

In this paper, we proposed a novel network for medical image segmentation, called DABU-Net. The DABU-Net takes advantage of the U-Net, the dilated convolution, and the attention mechanism. Firstly, we design the DCblocks with dilated convolution, and ASPP is used in our model, which makes DABU-Net can extract richer contextual information. Secondly, the spatial and channel attention are added to fill the semantic gap. Finally, we design a boundary augment module to enhance the relationship between segmentation region and boundary. The DABU-Net performances better when compared with other competitive methods on all three datasets. Moreover, we validate the effectiveness of proposed DCblock, attention module and boundary augment module, which makes inserting these modules into other networks as possible.

## References

1. Jonathan, L., Evan, S., Trevor, D.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2015, pp. 3431–3440. IEEE Computer Society, Boston (2015)
2. Olaf, R., Philipp, F., Thomas, B.: U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Munich (2015)
3. Nabil, I., M. Sohel, R.: MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **121**, 74–87 (2020)
4. Debesh, J., Michael, A.R., Dag, J., Pal, H., Havard, D.J.: DoubleU-Net: a deep convolutional neural network for medical image segmentation. In: CBMS 2020, pp. 558–564. IEEE, Rochester (2020)
5. Zaiwang, G., Jun, C., Huazhu, F., Kang, Z., Huaying, H., Yitian, Z., Tianyang, Z., Shenghua, G., Jiang, L.: CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Trans. Med. Imaging* **38**(10), 2281–2292 (2019)
6. Zongwei, Z., Md Mahfuzur Rahman, S., Nima, T., Jianming, L.: UNet++: a nested U-Net architecture for medical image segmentation. In: 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop 2018. Lecture Notes in Computer Science, vol. 11045, pp. 3–11. Springer, Granada (2018)
7. Ozan, O., Jo, S., Löic Le, F., Matthew, C.H.L., Mattias, P.H., Kazunari, M., Kensaku, M., Steven, G.M., Nils, Y.H., Bernhard, K., Ben, G., Daniel, R.: Attention U-net: learning where to look for the pancreas. CoRR, abs/1804.03999 (2018)
8. Reza, A., Maryam, A., Mahmood, F., Sergio, E.: Bi-directional ConvLSTM U-net with Densely connected convolutions. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, pp. 406–415. IEEE, Seoul (2019)
9. Dan, C.C., Alessandro, G., Luca Maria, G., Jürgen, S.: Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems, pp. 2852–2860. Nevada (2012)
10. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. CoRR, abs/1512.03385 (2015)
11. Gao, H., Zhuang, L., van der Laurens, M., Kilian, Q.W.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017, pp. 2261–2269. IEEE Computer Society, HI (2017)
12. Christian, S., Wei, L., Yangqing, J., Pierre, S., Scott, E.R., Dragomir, A., Dumitru, E., Vincent, V., Andrew, R.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2015, pp. 1–9. IEEE Computer Society, Boston (2015)
13. Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N.G., Lukasz, K., Illia, P.: Attention is all you need. In: Advances in Neural Information Processing Systems 30, Annual Conference on Neural Information Processing Systems 2017, pp. 5998–6008. California (2017)
14. Sehyung, L., Makiko, N., Hideyoshi, U., Haruo, K., Shin, I.: Mu-net: multi-scale U-net for two-photon microscopy image denoising and restoration. *Neural Netw.* **125**, 92–103 (2020)
15. Caiyong, W., Yong, H., Yunfan, L., Zhao Feng, H., Ran, H., Zhenan, S.: ScleraSegNet: an improved U-net model with attention for accurate sclera segmentation. In: 2019 International Conference on Biometrics, ICB 2019, Crete, Greece, pp. 1–8. IEEE, Crete (2019)
16. Abhijit, G.R., Nassir, N., Christian, W.: Concurrent spatial and channel ‘Squeeze & Excitation’ in fully convolutional networks. In: Medical Image Computing and Computer Assisted Intervention 2018. Lecture Notes in Computer Science, vol. 11070, pp. 421–429. Springer, Granada (2018)
17. Jie, H., Li, S., Gang, S.: Squeeze-and-excitation networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2018, pp. 7132–7141. IEEE Computer Society, Salt Lake City (2018)

18. Zhen-Liang, N., Gui-Bin, B., Xiao-Hu, Z., Zeng-Guang, H., XiaoLiang, X., Chen, W., Yan-Jie, Z., Rui-Qi, L., Zhen, L.: RAUnet: residual attention U-net for semantic segmentation of cataract surgical instruments. In: Neural Information Processing—26th International Conference, ICONIP 2019, Proceedings. Lecture Notes in Computer Science, vol. 11954, pp. 139–149. Springer, Sydney (2019)
19. Md. Zahangir, A., Mahmudul, H., Chris, Y., Tarek, M.T., Vijayan, K.A.: Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. CoRR, abs/1802.06955 (2018)
20. Liang-Chieh, C., George, P., Iasonas, K., Kevin, M., Alan, L.Y.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
21. Fisher, Y., Vladlen, K.: Multi-scale context aggregation by dilated convolutions. In: 4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings. San Juan (2016)
22. Mateusz, B., Ashirbani, S., Maciej, A.M.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. **109**, 218–225 (2019)

# Building Boundary Vectorization from Satellite Images Using Generative Adversarial Networks



Kunyue Yan , Yingxiao Xu , and Hao Chen

**Abstract** Building contours extracted by deep learning often have irregular boundaries, which need further regularization to get results more consistent with the actual building boundaries. In this paper, a regularization model of building boundary based on Generative Adversarial Networks is proposed. The irregular building contour generated by neural network is used as input, and the channel and spatial attention module are introduced to better learn the global information of the image. DP algorithm is used to refine the boundary of the learned building contour. The main direction of the building is get by finding the minimum circumscribed rectangle of the boundary. The boundary lines are adjusted in groups to get the orthogonal polygons. The experiment proved that compared with the extraction results of the R2U-Net, the method proposed in this paper can get more accurate and closer to actual building contours.

**Keywords** Building boundaries regularization · Satellite image · Generative adversarial networks

## 1 Introduction

With the rapid development of society, diverse and accurate geographic information plays an increasingly important role in urban construction and engineering application. As an important source of ground surface feature, high-resolution remote sensing image has rich spectral and texture information. Building is one of the important ground features, and it is of great significance to extract rapidly updated buildings from remote sensing images. Accurate building contour information has important applications in cartography, urban planning and so on. So far, there have been a great many of researches on building extraction from remote sensing images. However, those building contours extracted directly from remote sensing images are often irregular, which need to be processed later to get the vectorized boundary.

---

K. Yan · Y. Xu · H. Chen ()

National University of Defence Technology, Changsha 410000, China

e-mail: [hchen@nudt.edu.cn](mailto:hchen@nudt.edu.cn)

The methods of building extraction from remote sensing images are usually divided into traditional manual extraction methods and deep learning based methods. Traditional building extraction methods usually use the spectral features, texture features, morphological features of remote sensing image, or use some local features of remote sensing image buildings, such as corners, to match and extract. With the development of deep learning technology in recent years, building extraction from remote sensing image using deep learning methods has been a lot of research. Neural network can be used for pixel level segmentation or semantic level segmentation. The commonly used neural networks for image segmentation include CNN, U-Net, FCN, etc.

According to the analysis of the research status, this paper intends to use the deep learning network R2U-Net to extract the initial building mask from the remote sensing image, and then use the Generative Adversarial Networks [1] with the attention mechanism to generate a more regular building mask. By regularizing the direction of the building boundary, the vectorized boundary is obtained. The method in this paper can get more accurate and closer to actual building contours on the public data set INRIA.

## 2 Related Works

Building boundary vectorization can be divided into four methods: division based method, refinement based method, corner and line detection based method, and machine learning based method.

The method based on division usually classifies the satellite image initially according to the object to be extracted, and eliminates the interference caused by some useless information. After the initial classification, according to the spatial information and geometric characteristics of the building itself, the corresponding segmentation is carried out by using a specific appropriate algorithm. Sun et al. [2] first used the SVM algorithm to segment the image, and the building and non-building areas are obtained. After getting the masks of the building, the  $\alpha$ -extended algorithm and the energy function are constructed to segment and classify the building edge line. The advantage of this method is that the contour obtained by a series of operations is fine, and the visual effect is better. But the initial feature needs to be selected manually, and the operation is complex and the efficiency is not very high.

The refinement based method usually obtains the initial boundary points of buildings, and then adjusts the boundary points through a series of operations to obtain more regular building boundaries [3]. Douglas-Peucker algorithm [4] is a classical algorithm to remove redundant points. It has a good effect on the irregular boundary with serration. Zhao et al. [5] extract buildings from remote sensing images with Mask R-CNN, and then use Douglas-Peucker algorithm for initial refinement of boundary points. They select three points and the distance between them to build the model. By adjusting the position of the middle point, the sum of distances is calculated to find the optimal position of the second point, and then iterate in order to

get all points adjusted. Hong et al. [6] group LiDAR point cloud boundary data and then used Douglas-Peucker algorithm to remove the redundant points, screened the key points by judging the rationality of the boundary points, and finally adjusted the edge line by determining the main direction line of the building. The method based on refinement can get more regular building contour, but it needs more complete and high-quality building extraction results to have better performance.

Through corner detection of the extracted building mask, the key corners of the building contour can be obtained, and the regular building contour can be obtained by connecting the adjacent corners in a certain order [7]. JieXi et al. [8] compared the effect of corner extraction of buildings using Harris operator and Susan operator. Experimental analysis shows that Harris operator has better performance in corner detection. The method based on corner and line detection can obtain the building contour with geometric structure, but this method may be interfered by noise, shadow and other factors in the image, and cannot obtain the ideal results.

The method based on machine learning saves the tedious process of manual feature acquisition, and can learn all aspects of feature information [9, 10]. Inspired by Zorzi et al. [11, 12], this paper uses the GAN and combines line detection for refinement to get vectorized building boundaries.

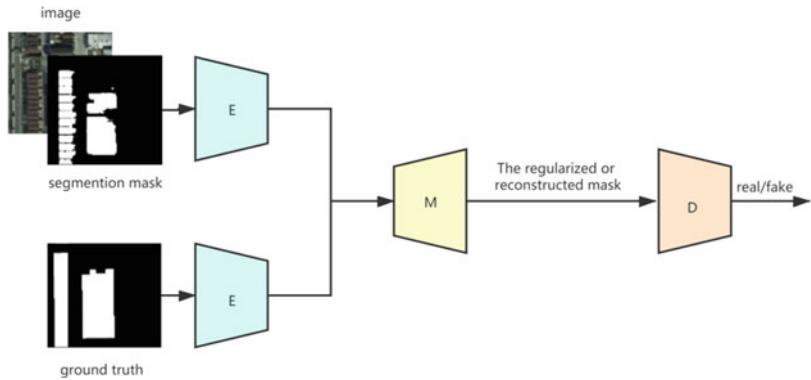
### 3 Methods

The basic method is to use R2U-Net to generate the initial irregular building mask, then regularize the mask by GAN with attention module, and finally generate the vectorized boundary by extracting the edge line and regularizing the boundary direction.

#### 3.1 *GAN with Attention Module for Regularization*

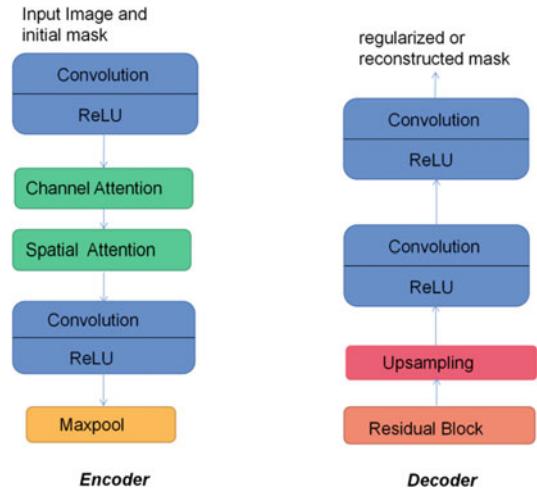
**Model Structure.** The model structure mainly includes generator G and discriminator D. the generator includes encoder-decoder structure, encoder E and decoder M. The building mask to be regularized and the corresponding image are used as the input of the generator to generate the regularized mask. Also, the ground truth mask is input into the encoder-decoder for reconstruction. The two are respectively input into the discriminator to determine whether the output is zero or one, so as to determine whether it is the reconstruction of the ground truth or the image generated by the decoder. The main structure is shown in Fig. 1.

The structure of the generator is a basic auto-encoder structure. It is showed in Fig. 2. The encoder consists of a series of  $3 \times 3$  convolution layer and the subsequent batch norm layer, max pool layer and ReLU layer as the activation function. After the convolution layer, the convolution block attention module (CBAM) [13] is



**Fig. 1** The structure of regularization model

**Fig. 2** The structure of the generator



added, which combines spatial attention and channel attention mechanism, so that the network can better learn the global content of the image.

The channel attention module takes the input feature map is processed by max pooling and average pooling and then by MLP. The MLP output features are added based on element wise, and then activated by sigmoid to generate the final channel attention feature map. The output characteristic graph of channel attention module is taken as the input characteristic graph of this module. It passes max pooling and global average pooling based on channel, and then does the concatenation of the two results based on channel. After a convolution operation, the dimension is reduced to one channel. Then the spatial attention feature is generated by sigmoid. Finally, the feature and the input feature of the module are multiplied to get the final feature.

The decoder consists of a series of residual layers,  $3 \times 3$  convolution layer, batchnorm layer and  $2 \times 2$  upsampling layer.

The discriminator has the same convolution block as encoder and decoder, but it has more max pool layer and has a sigmoid layer.

**Loss Function.**  $l_g$  is a conventional discriminative loss function of GAN. Its purpose is to update the parameters of generator G to make the mask generated by the model as close as possible to the real mask, and let discriminator D consider the regularized mask as a ground truth mask. The definition of binary cross entropy loss function is adopted, as shown in Eq. (1):

$$l_g = \text{bce\_loss}(D(G(a, b), \text{real})) \quad (1)$$

where  $a$  presents the input image and  $b$  presents the input irregular mask.

In the training process, LD, the loss function of discriminator D is also defined by the binary cross entropy loss function. The purpose of this is to update the parameters of discriminator to enable the discriminator to correctly distinguish the generated mask and the real mask. Its definition is shown in Eq. (2):

$$l_d = \text{bce\_loss}(D(c), \text{real}) + \text{bce\_loss}(D(G(a, b), \text{fake})) \quad (2)$$

where  $c$  is the reconstructed ground truth mask.

The binary cross entropy loss is used to calculate the loss of the generated mask and input mask, which makes generated mask close to the input mask. The loss of the reconstructed ideal mask and the input ideal mask is also calculated:

$$l_a = - \sum_i^N a_i \cdot \log G(a, b)_i \quad (3)$$

$$l_c = - \sum_i^N c_i \cdot \log G(c)_i \quad (4)$$

In order to make the generated regularized mask close to the input mask, the normalized cut loss and Potts loss [14, 15] is introduced:

$$l_n = \sum_k \frac{S^{kT} \hat{W} (1 - S^k)}{1^T \hat{W} S^k} \quad (5)$$

$$l_p = \sum_k S^{kT} W (1 - S^k) \quad (6)$$

where  $k$  is the label numbers and  $S$  is binary indicator vector.  $W$  is a matrix of discontinuity costs or affinity matrix.

The total loss function can be described as follows:

$$l = \alpha l_d + \beta l_a + \gamma l_c + \delta l_n + \varepsilon l_p \quad (7)$$

### 3.2 Boundary Direction Regularization

After getting the regularized mask, further optimization is carried out to obtain the vectorized boundary. After extracting the boundary point of the mask, the DP algorithm is used to remove the redundant boundary points to get the initial polygon. Then the main direction of the building is determined by finding the minimum circumscribed rectangle of the boundary point. According to the main direction, the direction of each side of polygon is adjusted to right angle polygon, which is more in line with the geometric characteristics of the actual building.

The specific adjustment methods are as follows:

- (1) Select a boundary point as the initial point, and judge the angle between the vector composed of it and the next point and the main direction vector.
- (2) If the included angle is less than the set threshold ( $45^\circ$ ), it is considered that the edge is parallel to the main direction, and the position of the next point is changed to make the edge conform to the parallel condition.
- (3) If the included angle is greater than the set threshold ( $45^\circ$ ), it is considered that the edge is perpendicular to the main direction, and the position of the next point is changed to make the edge conform to the vertical condition.
- (4) Iterate in sequence until each point is traversed.

## 4 Experiment Results

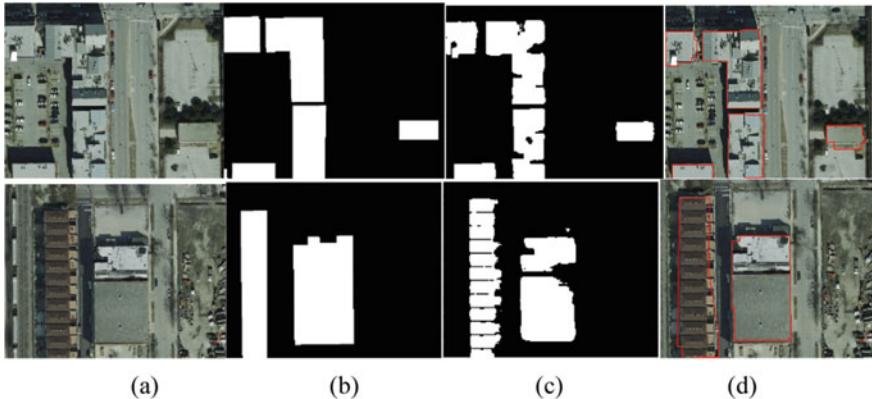
### 4.1 Dataset

The images from the public dataset INRIA are used, which contains 180 remote sensing images of multiple buildings. The size of each image is  $5000 \times 5000$ . In order to make the model training pay more attention to the details of the building, each image is cropped into 25 pieces of  $512 \times 512$  images for training. Choose three-fourths of them for training and the rest for testing.

### 4.2 Results

It can be seen from the figure that the building mask directly extracted by R2U-Net has irregular boundaries, and there are many gaps, small pieces in the masks. The extraction results are vulnerable to the interference of shadows. After boundary vectorization, a complete and geometric structure of the building boundary is obtained.

The results can also show that since the model has the attention model, compared with the directly extracted mask, the optimized image can pay more attention to



**Fig. 3** **a** is the initial image, **b** is the ground truth mask, **c** is the segmentation mask and **d** is the vectorized result

**Table 1** The evaluation scores of test area

Method	IoU	Precision	Recall
R2U-Net	0.441	0.952	0.457
Ours	0.447	0.904	0.462

the global distribution, so that the original scattered buildings can have the overall boundary after optimization.

Because the network learning and subsequent optimization are carried out on the initial building extraction results, the effect of this method depends on the quality of the initial extraction results. If the initial results are poor, it is likely that the ideal effect will not be achieved (Fig. 3).

The specific evaluation scores are shown in the Table 1. It seems that ours IoU and the Recall scores are slightly higher than R2U-Net results. The precision score is lower than that of R2U-Net. Since the regularization process is carried on the basis of the segmentation result, it is influenced by the results of the segmentation quality.

## 5 Conclusion

On the basis of building information extraction by deep learning, this paper proposes an idea of building boundary vectorization. For the problems of jagged and irregular boundary in building image extraction by deep learning, a regularization method based on GAN is proposed. Then the redundant points of building contour are removed by the DP algorithm to get the initial contour polygon, and the main direction is determined by selecting the minimum circumscribed rectangle. Thus, the direction of polygon line is regularized, and the actual building boundary contour is obtained. Through the experimental verification, the method introduced in this paper

in the deep learning extraction of high-resolution remote sensing image makes the extracted results more closer to the actual image of the building boundary.

## References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
2. Shaoguang, Z., Jinyan, S., Li, F., Jing, X., Chao, C.: Extraction of building contour from high resolution images. *Remote Sens. Land Resour.* **27**(3), 54–58 (2015)
3. Sohn, G., Jwa, Y., Jung, J., Kim, H.: An implicit regularization for 3D building rooftop modeling using airborne lidar data. *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.* I-3(September), 305–310 (2012)
4. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: Int. J. Geogr. Inf. Geovisualization* **10**(2), 112–122 (1973)
5. Zhao, K., Kang, J., Jung, J., Sohn, G.: Building extraction from satellite images using mask RCNN with building boundary regularization. In: CVPR Workshops, pp. 247–251 (2018)
6. Shaoxuan, H., Feng, Y., Jingxue, W.: Research on building boundaries regularization algorithm for LiDAR point clouds. *Sci. Surv. Mapp.* **45**(07), (2020)
7. Zhou, Z., Wang, J., Zhu, Q., Liu, X., Ma, Z., Gao, X.: Remote sensing image building contour optimization method based on minimum external rectangle. *Beijing Surv. Mapp.* **35**(1), 1–6 (2021)
8. Jiexi, W., Dejun, F.: A building boundary regularization method by contrasting Harris operator and Susan operator. *Bull. Surv. Mapp.* **04**, 11–15 (2020)
9. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: High-resolution aerial image labeling with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **55**(12), 7092–7103 (2017)
10. Wei, S., Ji, S., Lu, M.: Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Trans. Geosci. Remote Sens.* **58**(3), 2178–2189 (2020)
11. Zorzi, S., Fraundorfer, F.: Regularization of building boundaries in satellite images using adversarial and regularized losses. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2019)
12. Zorzi, S., Bittner, K., Fraundorfer, F.: Machine-learned regularization and polygonization of building segmentation masks. In: IEEE International Conference of Pattern Recognition (2020)
13. Woo, S., Park, J., Lee, J., Kweon, I.: CBAM: convolutional block attention module. In: CVPR Workshops, pp. 3–19 (2018)
14. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised CNN segmentation. arXiv preprint [arXiv:1803.09569](https://arxiv.org/abs/1803.09569) (2018)
15. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised CNN segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City (2018)

# Research on Tomato Maturity Detection Based on Machine Vision



Sen Lian , Linlin Li , Weibin Tan , and Lixin Tan

**Abstract** Maturity as one of the important criteria for identifying tomato quality, at the same time, it is also the main basis for automated grading in industry. To prevent data collection from being affected by other interference factors and improve the standardization of tomato data collection, a parallel-angle single-sided tomato image collection system is designed for the problem of different colors of tomatoes and difficulty in manual grading. Under the HSV color theory, a custom threshold segmentation method based on H component is proposed. This method determines that  $0\text{--}8^\circ$  and  $156\text{--}180^\circ$  are the best hue thresholds for pink tomato varieties in the HSV color space. After the morphological calculation process, the ratio of the binarized red pixels to the tomato outline pixels is used to calculate the maturity level, which verifies that the tomato grades of different maturity quality can be effectively distinguished.

**Keywords** Machine vision · Tomato grading · HSV color space · Maturity

## 1 Introduction

Tomatoes are not only rich in nutritional value, but also have a wide range of uses. They can be used to make a variety of processed products, such as tomato juice, peeled tomatoes, etc., and are deeply loved by people. China is a big tomato producer in the world. The quality inspection and grading of fruit and vegetable products have always been a huge and complex task. Individual grading cannot maintain a constant standard. It is inevitable that there will be detection errors, which will affect work efficiency and grading accuracy. Surface color is an important basis for the grading of external quality of tomatoes [1], but relying on manual tomato grading is very cumbersome. Human eyes cannot accurately determine the diameter, color, shape,

---

S. Lian · L. Li · W. Tan  
Hunan Agricultural University, Changsha, China

L. Tan   
Hunan College of Information, Hunan Agricultural University, Changsha, China  
e-mail: [tanlixin@mail.hnu.cn](mailto:tanlixin@mail.hnu.cn)

maturity, etc. The quality of the sorted fruits is uneven, which brings difficulties to subsequent packaging, storage and transportation, processing and sales.

The technology of using hyperspectral, multispectral or infrared to analyze crop images has been widely used, and while the effect is remarkable, there are also problems such as high cost and difficult operation. Syahrir et al. [2], converted the RGB chromaticity space into the lab chromaticity space, and after preprocessing the image, the ripeness of the tomato was judged by the R-G chromatic aberration. Wan et al. [3], segmented the color area in tomatoes with the maximum circle cutting method, Input the vector mean value of chromaticity under RGB and HIS theory into BP neural network for maturity research. Wang [4] transformed the tomato image into the HIS color model for maturity grading, but it has limitations under the influence of light. Huang [5] uses an improved canny edge detection algorithm to improve the effect of light noise on the apple image and improve the effect of contour extraction. Sun [6] transformed the remote sensing image in the visible light band into the HSV color space, which enhanced the contrast of the target cloud area. Yang [7] judges the maturity of citrus by analyzing the ratio of yellow and green pixels. Dangdi [8] separated the HSV three-channel combined with the OTSU segmentation algorithm to identify the shooting target surface. Hou [9] carried out experiments on garlic buds with bilateral image recognition. Xia [10] uses K-means clustering and OTSU to segment cotton under the HSV color model, and the accuracy can reach 80%. Chen [11] uses machine vision to perform defect detection, size and color grading of dragon fruit. This topic is based on machine vision and uses high-definition cameras to collect double-sided data on tomatoes, and then use python to process the synthesized image data to obtain the color, size, shape and surface defects that determine the quality of the tomato at one time. This information is used for tomato detection and classification, which has the advantages of fast speed, high accuracy, and no loss of tomato quality.

Tomato color is an important basis for grading maturity. This topic uses HSV color space theory and machine vision algorithm to design a set of tomato image acquisition platforms, researches the threshold of tomato maturity under the H component, the color thresholds of different maturity tomatoes are studied experimentally, and the maturity level of tomatoes is judged by the ratio of the color extracted from the feature and the overall outline pixel after the binarization process.

## 2 Tomato Image Acquisition Platform

### 2.1 Build Image Acquisition Platform

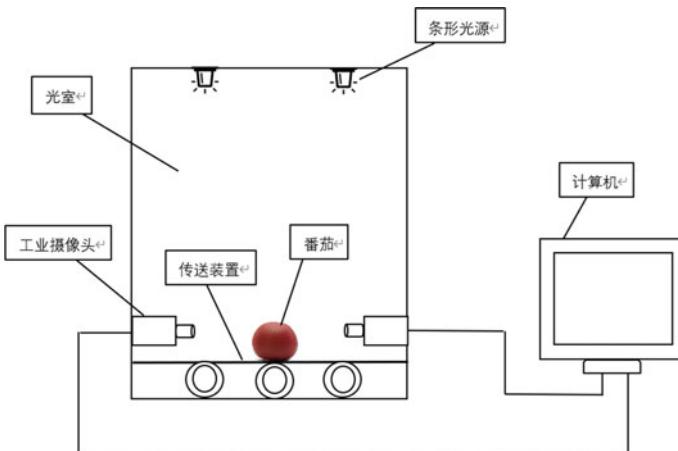
Ensure that the image collection process is not affected by other interference factors. In this project, a set of tomato data collection platforms is built to keep the external factors consistent during data collection. This platform is in a closed dark box. Considering the influence of light shadow and tomato surface reflection on the

imaging quality, the matte board is selected as the base for carrying the tomato, and the LED strip light source is selected to provide light to ensure the uniform divergence of the light. In order to ensure the clarity and texture characteristics of the collected data, an industrial camera was selected as the information collection device, and the ring LED light source was selected to increase the image quality to increase the exposure of the industrial camera. Since the tomato data is collected from both sides, two industrial cameras are equipped. The data is processed by the computer. The specific standards are as follows:

- (1) The brightness of the light source in the data collection light room is kept consistent;
- (2) The setting of the exposure parameter value of the industrial lens LED ring light source is consistent;
- (3) When collecting, the tomato is placed in the calibration position of the information collection platform to ensure that the distance between the data collected every time is the same;
- (4) Two sets of industrial cameras with ring light sources are installed at a fixed location, the shooting distance is fixed, the shooting angle is fixed and parallel to the collection object, and the camera shooting parameters are kept consistent.
- (5) Choose a center position of the platform to shoot on both sides at an angle parallel to the object to be collected, and fix the camera at a position of 20 cm in the center.

The hardware structure diagram of the data acquisition platform designed by this subject is shown in Fig. 1.

There are also separate links to the user guide, which can be referred to by the user.



**Fig. 1** Schematic diagram of data acquisition platform structure

## 2.2 Camera Installation Location

The fruit grading of agricultural products has increasingly become one of the directions of machine vision applications. Through studying the literature of a large number of scholars, it is found that most of the experimental subjects' information collection methods are taken from a bird's-eye view angle. This angle data is conducive to shape and automatic grading of dimensions. Considering that the full picture of the collected subjects cannot be presented well, it is not conducive to the color analysis in the later stage. Therefore, this subject chooses to take single-sided shooting at an angle parallel to the collection object, and synthesize images on the front and back of the same collection object. In order to prevent image distortion, the synthesized picture adopts a 1:1 ratio without any scaling. After experiment and comparative analysis, fix the camera at the center of 20 cm away from the centroid of the object to be collected, so this distance is chosen to fix the position of the industrial camera.

## 2.3 Image Acquisition Object

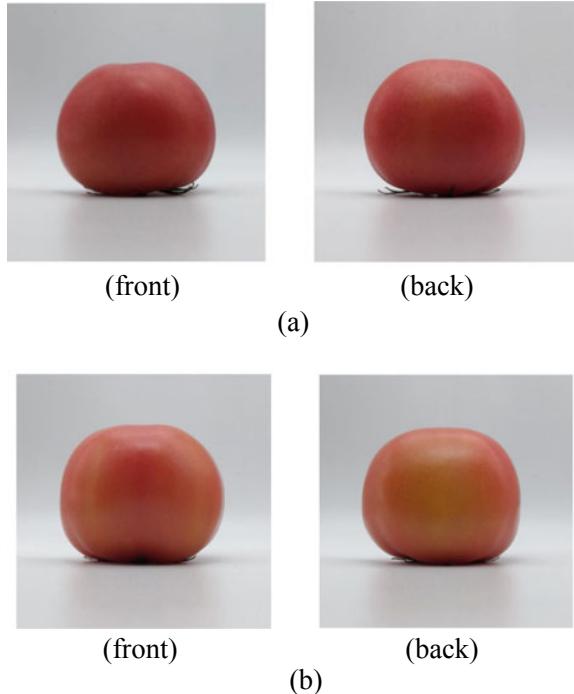
The research content of this topic is how to detect tomato maturity through machine vision. Therefore, pink tomatoes in tomato species are selected as the information collection object of this experiment, and white is selected as the background to reduce the influence of reflection on the bottom plate. When the image is collected, the tomato is placed in the calibration position of the information collection platform. Part of the tomato collected images is shown in Fig. 2. By comparing Fig. 2a and b, we can find that there are subtle differences between the two images. Figure 2a shows that the tomato has a uniform color and a darker maturity. Picture 2b shows that part of the tomato is greenish green. Therefore, determining the appropriate threshold is very important for the later stage of computing maturity.

## 3 Color Model

### 3.1 RGB Color Space

RGB is a commonly used way of expressing color information. It is designed from the principle of color luminescence. The tomato data collected in this topic is also stored in RGB format I. This theoretical combination superimposes different primary colors and transforms them into the required colors. It is a more commonly used color format, and the standard value range is between 0 and 255. Its basic primary colors are red, green, and blue. It is represented by points in the three-dimensional space. The changes of the three primary colors can combine various colors, and the brightness

**Fig. 2** Original image of tomato



will also affect their components. There are three pairs of complementary colors in these colors, namely red and cyan, purple and green, yellow and blue, the value range of the three primary colors is R: 0–255; G: 0–255; B: 0–255, the value can be normalized to 0–1 after dividing by 255[12]. Since human vision has different sensitivity to the three components of R, G, and B, if the degree of color similarity is measured by Euclidean distance, the result will have a large error with vision [8].

### 3.2 HSV Color Space

HSV color theory proposes a color model of hexagonal cone, which is a model for human eye color perception. In the HSV color space, the information content is represented by three attributes Hue (H), the value range is: 0°–360°, the three complementary colors are red 0° and yellow 60°, green 120° and cyan 180°, blue 240° and purple 300°. Saturation (S), the value range is 0 to 100%. Brightness (V), the value range is 0% (black) to 100% (white) [12]. Hue is related to the wavelength of the main light in the mixed spectrum. Different wavelengths of light show different colors and also reflect the difference in color tone [8]. Because HSV is more suitable for capturing objects with brighter colors.

### 3.3 Conversion Principle of RGB and HSV

The RGB color model is based on the three-dimensional coordinates established by the Cartesian coordinate system. The r, g, and b channels are located on the three-dimensional coordinate system, with red, green, and blue as the primary colors. The central axis from the origin to the white vertex is the gray line  $rgb$  is equal. The HSV color model converts the three-dimensional coordinate system of the RGB color model into a cone-shaped subset, and the value of the vertex V of the cone subset is 1. It contains the three faces of R = 1, G = 1, and B = 1 in the RGB model. The hue H is around the brightness V axis. Rotate 360° to form a circle, the saturation S is the proportional value, and the value range is [0, 1]. The conversion relationship is as follows:

$$V = \max(rgb) \quad (1)$$

$$S = (\max(rgb) - \min(rgb)) / \max(rgb) \quad (2)$$

$$\max = \min, H = 0^\circ \quad (3)$$

If the maximum value is r and g more than the b:

$$H = 60^\circ * (g - b) / (\max - \min) + 0^\circ \quad (4)$$

If the maximum value is r and g Less than b:

$$H = 60^\circ * (g - b) / (\max - \min) + 360^\circ \quad (5)$$

If the maximum value is g:

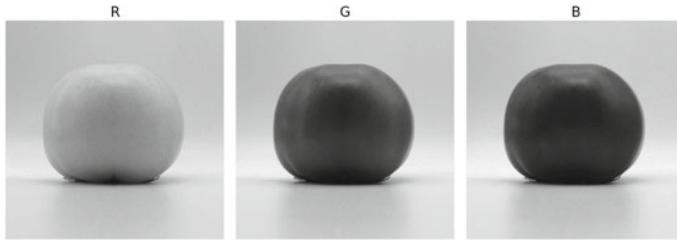
$$H = 60^\circ * (b - r) / (\max - \min) + 120^\circ \quad (6)$$

If the maximum value is b:

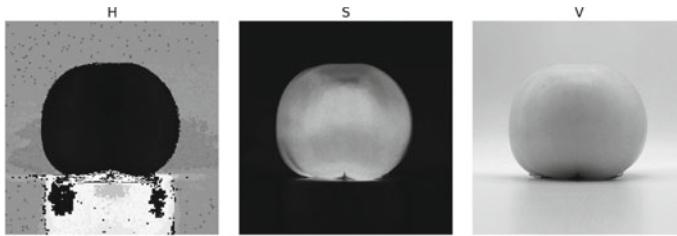
$$H = 60^\circ * (r - g) / (\max - \min) + 240^\circ \quad (7)$$

### 3.4 Comparative Analysis of Color Components

Tomatoes are mainly bright red, and the surface of underripe or immature tomatoes has turquoise. In order to better choose which color model is more helpful for extracting the color of the tomato surface, this topic is based on the RGB color space,



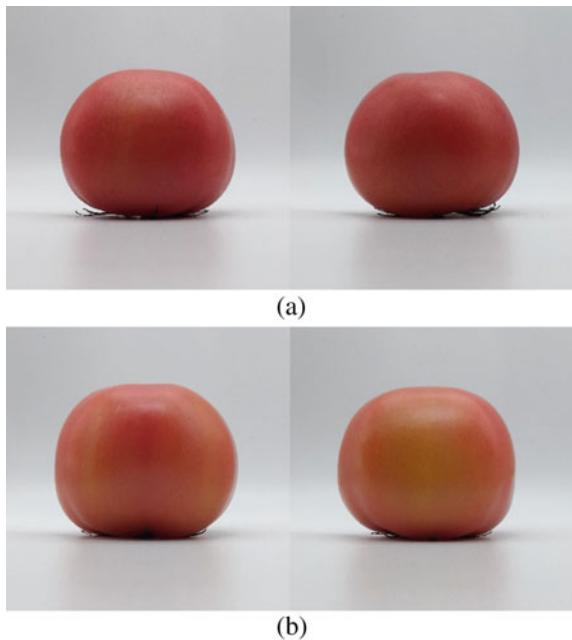
**Fig. 3** Three-channel grayscale image of RGB color space



**Fig. 4** Three-channel grayscale image of HSV color space

the three-channel R component, G component, B component compare the grayscale images of tomato find that the difference is not big. Analyze the grayscale image of tomato under the H component, S component and V component of the three channels of HSV color space, the difference is obvious. Take Fig. 2(b front side) as the data, Take the data as an example to observe the pixel change intensity of each channel. The three-channel grayscale image of RGB color space is shown in Fig. 3, and the three-channel grayscale image of HSV color space is shown in Fig. 4.

Comparing and analyzing Figs. 3 and 4, it can be seen that the tomato characteristics based on the RGB color space have not changed significantly, and the brightness of the R channel has a large change. Tomato feature changes based on HSV color space are more obvious, especially the H tone channel, which can well show the reflection and noise pollution when collecting tomato images. By comparing the original image of Fig. 2(b front), it can be found that on both sides of the tomato picture in Fig. 2(b front) there are fine vertical strips of turquoise, which contrast with the red that occupies most of the tomatoes, which can be reflected in the reflective imaging.. It can be seen from the above experiments that the HSV color model is more sensitive to changes in tomato surface color, and is more suitable for the research of tomato maturity analysis in this subject.

**Fig. 5** 1:1 composite picture

## 4 Judgment of Tomato Maturity Grad

### 4.1 Image Preprocessing

Cut the tomato images collected by the two cameras to a specified size of  $2000 \times 2000$  pixels, and combine the same tomato with a 1:1 ratio losslessly into a picture, as shown in Fig. 5.

The goal of this subject is to recognize tomatoes and pay more attention to the reduction of image impurities. Therefore, this subject uses Gaussian filtering to process the image to lay the foundation for finding the best threshold.

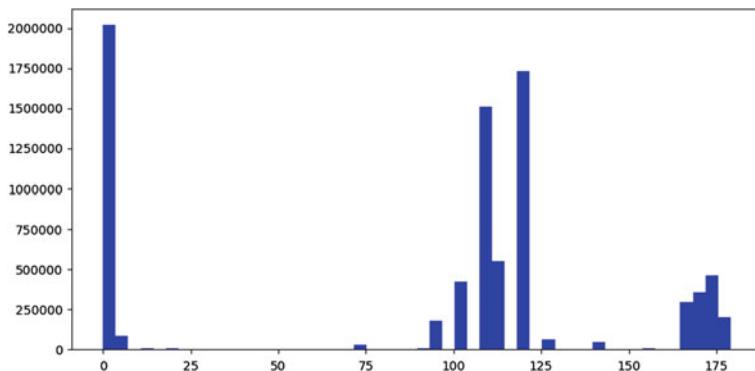
### 4.2 Image Segmentation

In OpenCV, the value range of H is  $0\text{--}180^\circ$ , and both S and V are  $0\text{--}255$ . According to Table 1, the colors corresponding to the HSV component range can be known, as shown in the following table:

It can also be seen from the hue channel histogram of tomato in the HSV color space in Fig. 6 that the denser H channel components of the histogram are mainly concentrated around  $0^\circ$  and  $175^\circ$ , and most of the rest are occupied by the white background.

**Table 1** HSV color value comparison table

	Black	White	Red1	Red2	Green	Cyan	Blue
hmin	0	0	0	156	35	78	100
hmax	180	180	10	180	77	99	124
smin	0	0	43		43	43	43
smax	255	30	255		255	255	255
vmin	0	221	46		46	46	46
vmax	46	255	255		255	255	255

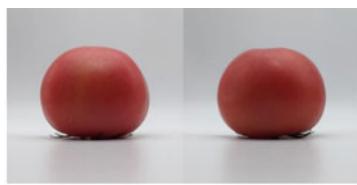
**Fig. 6** Tomato H channel color histogram

After the synthesized tomatoes of different maturity levels are processed by Gaussian filtering, refer to the red hue component threshold range shown in Table 1, and then combine the H-component histogram distribution range of the red tomato in Fig. 6 to extract the color of the custom threshold and pass Morphological operation. After a lot of experimental analysis, it is determined that the optimal threshold of tomato under the HSV color model is  $(0-8, 60-255, 60-255) \cup (156-180, 60-255, 60-255)$ , and the core size is 5 corroded Expansion operation 6 times achieves the best segmentation effect, which can effectively distinguish the mature and under-ripe parts of tomato. The experimental result comparison chart is shown in Fig. 7.

#### 4.3 Judgment of Tomato Maturity Grade

The calculation of tomato maturity is measured by the percentage of white pixels  $w_{(i,j)}$  in the tomato binarization map to the overall contour pixels  $s_{(i,j)}$  of the tomato. The calculation formula is:

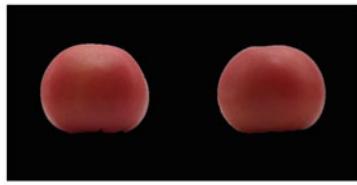
**Fig. 7** Experimental results comparison chart



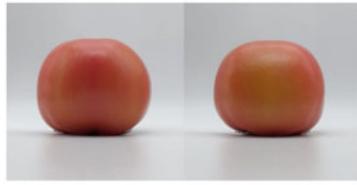
(a) Original image of ripe tomatoes



(a) Ripe Tomato Binarization Diagram



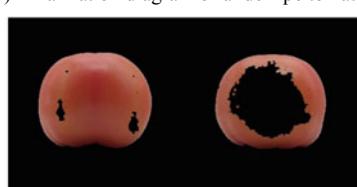
(a) Ripe tomato mask illustration



(b) Original image of underripe tomatoes



(b) Binarization diagram of underripe tomatoes



(b) Underripe Tomato Mask Illustration

**Table 2** Tomato ripeness grade standard

Proportion of red pixels	Maturity level
$S > 96\%$	Premium
$85\% < S < 96\%$	Level 1
$75\% < S < 85\%$	Level 2
$S < 75\%$	Immature

$$S = \sum_0^n w(i, j) / s(i, j) * 100\% \quad (8)$$

This subject has customized a set of tomato maturity grading standards, as shown in Table 2.

In order to verify the accuracy of the model, this experiment picks 100 tomatoes that are about to reach the growth cycle from the agricultural greenhouse as a sample library, and randomly selects 20 tomatoes as the experimental data at an interval of 2 days. After the test is completed, 20 tomato samples are selected. Put it back into the sample library, and continue to select randomly in the next cycle. The data for manually judging tomato standards in each cycle is shown in Table 3.

Every cycle, 20 tomato samples are put into the model prepared in this topic and verified by computer. The verification results are shown in Table 4.

Combining the results of the verification of the accuracy of tomato maturity in Table 4, the judging of extra-grade and immature tomatoes is basically correct, and the judging results of the first and second maturity tomatoes are not good. The main source of error is the second level of tomato maturity and the first level of maturity. The boundaries between the levels are not obvious. Taking into account the subtle differences between computer calculation errors and manual judgments, this result shows that the use of this model can replace manual analysis of tomato maturity.

**Table 3** Tomato sample data table

Period (days)	Immature	Level 2	Level 1	Premium
First cycle	14	4	2	0
Second cycle	9	8	3	0
Third cycle	1	8	8	3
Fourth cycle	0	3	11	6

**Table 4** Accuracy rate of tomato maturity verification

Period (days)	Accuracy (%)	Source of error
First cycle	85.0	Immature 2pcs, Level 2 1pcs
Second cycle	90.0	Immature 1pcs, Level 2 1pcs
Third cycle	85.0	Level 2 2pcs, Level 1 1pcs
Fourth cycle	80.0	Premium 1pcs, Level 1 3pcs

## 5 Conclusion

Machine vision has strong applicability for tomato maturity grading algorithm. The tomato image acquisition platform designed in this subject can reduce the interference of light reflection and shadow on tomato imaging quality and fully collect tomato characteristics. Calculate the ratio of the pixels of the extracted color features to the tomato contour pixels through the segmented tomato binary image. According to the tomato maturity grade standard, it can replace the manual tomato maturity analysis.

However, because the computer is processing the HSV color model image of the tomato, it is limited and the performance of the computer is slightly insufficient in processing speed, the tomato needs to stay on the conveyor belt for 2–3 s to provide sufficient time to calculate the data when collecting data. In addition, this experiment did not consider the scenario where multiple targets are recognized at the same time, therefore, has certain limitations, and the algorithm needs to be improved.

**Acknowledgements** Professor Li-xin Tan. Lin-lin Li. Wei-bin Tan.

## References

1. Feng, B.: Computer vision classification of fruit based on fractal color. *Trans. CSAE* **18**(2), 141–144 (2002)
2. Syahrir, W.M., Suryanti, A., Connyngham, C.: Color grading in tomato maturity estimator using image processing technique. In: Li, W., Zhou, J. (eds.) *Proceedings of 2009 2nd IEEE International Conference on Computer Science and Information Technology*, vol. 2, pp. 290–294. Computer Science and Information Technology (2009)
3. Wan, P., Toudeshki, A., Tan, H., et al.: A methodology for fresh tomato maturity detection using computer vision. *Comput. Electron. Agric.* **146**, 43–50 (2018)
4. Wang, W., Yongjie, C.: Research on tomato color detection based on computer vision (01), 49–51 (2017). CNKI:SUN:NJTU.0
5. Huang, C., Fei, J.: Apple online grading method based on image feature fusion. *Trans. Chin. Soc. Agric. Eng.* **33**(1), 285–291 (2017)
6. Sun, H.: Remote sensing image fast cloud detection based on HSV color space. *Geospatial Inf.* **18**(08), 35–40 (2020)
7. Yang, H.: Research on Feature Extraction and Classification Algorithms of Typical Products Based on Machine Vision. <https://kns.cnki.net/KCMS/detail.aspx?dbname=CMFD201801&filename=1018071835.nh> (2017)
8. Di, D., Peifeng, Z.: Shooting target surface recognition based on HSV color space and OTSU algorithm. *Intell. Comput. Appl.* **10**(07), 11–16 22 (2020)
9. Hou, J.: Design and experiment of garlic sprout and seeding test bench based on bilateral image recognition. *Trans. CSAE* **36**(01), 50–58 (2020)
10. Xia, Y.: Cotton HSV image segmentation based on K-means clustering and two-dimensional Otsu. *Software* **41**(07), 170–173 (2020)
11. Chen, L.: Design of automatic grading system for dragon fruit based on machine vision. *Agric. Mech. Res.* **42**(05), 130–133 (2020)
12. Mengxia, H.: Color image segmentation based on RGB color space. *Comput. Knowl. Technol.* **16**(34), 225–227 (2020)

# Correlation Filter RGB-T Tracker with Modality and Channel Reliability



Fei Zhang and Shiping Ma

**Abstract** RGB-T object tracking is developing rapidly in the past decade due to the complementarity of visible (RGB) and thermal infrared (T) images. However, many trackers using multi-modal information by simple feature concatenation, which ignores both the modality and channel reliability. In this paper, we propose a correlation filter-based RGB-T tracker to learn the reliability weights in terms of modality and inter-channel. Specifically, the channel regularization collaborates with the spatial regularization to jointly learn the filter and channel weights. Besides, we design a novel objective function to optimize the modality reliability weight frame by frame. Through the reliability evaluation, the useful information hidden in the modalities and channels is fully exploited. We perform extensive experiments on the RGB-T benchmark, i.e., GTOT, to verify the effectiveness of the proposed method. Experimental results show that the proposed fusion strategy can improve tracking performance.

**Keywords** RGB-T tracking · Correlation filter · Adaptive fusion · Channel attention

## 1 Introduction

Visual object tracking is one of the fundamental tasks in computer vision and image process. Generic object tracking performs tracking based on visible images. With the development of sensor technology, tracking with both visible and infrared images (RGB-T tracking) has received more and more attention.

Although visible images have rich color and texture information, tracking task is difficult to function in the conditions of strong light or weak light. Fortunately, infrared images are not sensitive to light illumination. Therefore, both visible and

---

F. Zhang · S. Ma ()  
Air Force Engineering University, Xi'an 710038, China  
e-mail: [mashiping@126.com](mailto:mashiping@126.com)

infrared images provide a new opportunity to advance tracking performance in challenging scenarios, such as illumination variation and thermal crossover. However, how to effectively fuse different modalities is still an urgent issue to be solved.

Recently, some works focus on RGB-T fusion tracking. According to different fusion methods, these methods can be categorized as image-level [1, 2], feature-level [3], and decision-level [4, 5]. For comparison, Li et al. [4] extend some RGB trackers (including correlation filter trackers and deep trackers) to RGB-T trackers by directly feature concatenation. However, the fusion method only considers the collaboration between visible and infrared modalities while neglects the discrepancy of different modalities. Besides, most existing trackers ignore the use of feature channel reliability, which can reflect the contribution of different channels.

To address the above problems, we propose a real-time correlation filter RGB-T tracker via both modality and channel reliability evaluation, named MCCF. Specifically, the channel regularization is integrated into the objective function to jointly learning the filter and reliability of feature channel, which is aimed at full use of the feature in the channel dimension. Furthermore, we propose an adaptive fusion strategy to learn a reliability weight of each modality. The proposed MCCF can be effectively optimized by the ADMM algorithm. Experiments on GTOT [6] benchmark demonstrate that MCCF can achieve promising tracking performance while running at real-time speed.

The main contributions of this paper can be summarized as follows:

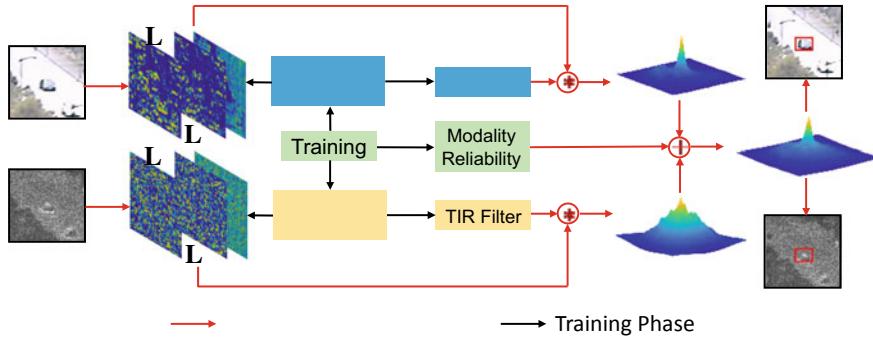
- (1) We propose a joint spatial-channel regularization to fully mine the power of feature in terms of channel.
- (2) An additional objective function is designed to learn the reliability weight of each modality.
- (3) Extensive experiments on GTOT benchmark shows that the proposed tracker has comparable performance against other state-of-the-art RGB and RGB-T trackers.

## 2 The Proposed Method

In this section, we propose a new correlation filter RGB-T tracker with reliability evaluation of both modalities and channels. Figure 1 shows the pipeline of the proposed MCCF tracker.

### 2.1 BACF

The classic CF-based tracker BACF is selected as our baseline tracker. Given the feature of the input image  $x \in R^{N \times 1 \times C}$ , the desired filter  $w \in R^{N \times 1 \times C}$  can be obtained by the following function:



**Fig. 1** The tracking framework of the proposed tracker. In the training phase, a unified loss is optimized to obtain both the RGB and TIR filters. And an additional objective function can be directly solved to learn a reliability weight of each modality. In the detection phase, channel and modality reliability weights are used for adaptive fusion based on decision-level (response-level)

$$L(w) = \frac{1}{2} \left\| y - \sum_{c=1}^C P^T x^c * w^c \right\|_2^2 + \frac{\lambda}{2} \sum_{c=1}^C \|w^c\|_2^2 \quad (1)$$

where  $P \in \mathbf{R}^{M \times N}$  ( $M = N$ ) is used to crop more true negative samples and  $T$  is the transpose operation. The first term is the regression term for regressing the filter and the second term is regularization term for avoiding over-fitting.

## 2.2 Channel-Spatial Regularized Correlation Filter

**Overall Function.** In order to fully exploit the feature from each channel, we propose a novel adaptive channel-aware correlation filter tracking method. Based on BACF, the overall function of the proposed method can be expressed as follows,

$$L(w, \beta) = \frac{1}{2} \left\| y - \sum_{c=1}^C \beta^c P^T x^c * w^c \right\|_2^2 + \frac{\lambda}{2} \sum_{c=1}^C \|w^c\|_2^2 + \frac{\gamma}{2} \|\beta - \beta^r\|_2^2 \quad (2)$$

where  $\beta = [\beta^1, \beta^2, \dots, \beta^C] \in \mathbf{R}^C$  denotes the importance weight of each channel and  $\beta^r \in \mathbf{R}^C$  stands for the reference channel weights. Thus, the filter and the channel weight can be optimized jointly to learn a more robust model for tracking.

**Optimization.** Denoting the auxiliary variable  $u^c = P^T w^c$ , the Fourier form of Eq. (2) can be described as,

$$L(w, \hat{u}, \beta) = \frac{1}{2N} \left\| \hat{y} - \beta^T \hat{X} \hat{u} \right\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 + \frac{\gamma}{2} \|\beta - \beta^r\|_2^2 \quad (3)$$

where the superscript  $\hat{\cdot}$  represents the Discrete Fourier Transform. Then, the Augmented Lagrangian Method is applied to Eq. (3).

$$L(w, \hat{u}, \beta) = \frac{1}{2N} \left\| \hat{y} - \beta^T \hat{X} \hat{u} \right\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 + \frac{\gamma}{2} \left\| \beta - \beta^r \right\|_2^2 + \frac{\mu}{2} \left\| \hat{u} - \sqrt{N}(FP^T \otimes I_D)w \right\|_2^2 \quad (4)$$

where  $\otimes$  denotes the Kronecker product and  $\mathbf{F}$  represents the orthonormal matrix. Here, Eq. (4) can be transformed into three subproblems to obtain their own closed-form solutions.

*Subproblem w.*

$$w^* = \frac{\lambda}{2} \|w\|_2^2 + \hat{\zeta}^T (\hat{u} - \sqrt{N}(FP^T \otimes I_D)w) + \frac{\mu}{2} \left\| \hat{u} - \sqrt{N}(FP^T \otimes I_D)w \right\|_2^2 \quad (5)$$

The solution of subproblem **w** can be acquired in the temporal domain,

$$w^* = \frac{\mu u + \zeta}{\mu + \lambda/N} \quad (6)$$

*Subproblem u.*

$$\begin{aligned} L(\hat{u}) &= \frac{1}{2N} \left\| \hat{y} - \beta^T \hat{X} \hat{u} \right\|_2^2 + \hat{\zeta}^T (\hat{u} - \sqrt{N}(FP^T \otimes I_D)w) \\ &\quad + \frac{\mu}{2} \left\| \hat{u} - \sqrt{N}(FP^T \otimes I_D)w \right\|_2^2 \end{aligned} \quad (7)$$

Equation (7) can be decomposed into N smaller problems,

$$\begin{aligned} \hat{u}(n) &= \frac{1}{2N} \left\| \hat{y}(n) - \hat{x}(n)^T B \hat{u}(n) \right\|_2^2 + \hat{\zeta}(n)^T (\hat{u}(n) - \hat{w}(n)) \\ &\quad + \frac{\mu}{2} \left\| \hat{u}(n) - \hat{w}(n) \right\|_2^2 \end{aligned} \quad (8)$$

Taking the derivative of Eq. (8) and making the outcome zero, we can get

$$\hat{u}(n) = \frac{B \hat{x}(n) \hat{y}(n) + \mu N \hat{w}(n) - N \hat{\zeta}(n)}{B \hat{x}(n) \hat{x}(n)^T B + \mu N I_D} \quad (9)$$

where  $B = \text{diag}(\beta)$ . The Sherman-Morrison formula is used to decrease the computational complexity of Eq. (9),

$$\begin{aligned}\hat{u}(n) &= \frac{1}{\mu N} (B\hat{x}(n)\hat{y}(n) + \mu N\hat{w}(n) - N\hat{\zeta}(n)) \\ &\quad - \frac{B^2\hat{x}(n)\hat{x}(n)^T}{\mu N E} (B\hat{x}(n)\hat{y}(n) + \mu N\hat{w}(n) - N\hat{\zeta}(n))\end{aligned}\tag{10}$$

where  $E = \mu N + \hat{x}(n)^T B^2 \hat{x}(n)$ .

*Subproblem  $\beta$ .*

$$L(\beta) = \frac{1}{2N} \left\| \hat{y} - \beta^T \hat{X} \hat{u} \right\|_2^2 + \frac{\gamma}{2} \left\| \beta - \beta^r \right\|_2^2\tag{11}$$

For convenience, abbreviate  $\hat{X}\hat{u}$  as  $\hat{M}$ . The solution of  $\beta$  can be directly obtained by setting the derivative about  $\beta$  to zero:

$$\beta = \frac{\hat{M}^T \hat{y} + \mu N \beta^r}{\hat{M}^T \hat{M} + \mu N}\tag{12}$$

*Subproblem  $\hat{\xi}$ .*

$$\hat{\xi}^{(i+1)} = \hat{\xi}^{(i)} + \mu(\hat{u}(n) - \hat{w}(n))\tag{13}$$

where  $\mu$  is updated as  $\mu^{(i+1)} = \min(\mu_{\max}, \delta\mu^{(i)})$ .

## 2.3 Adaptive Modality Fusion

Most of the existing CF-based trackers directly concatenate feature vectors of both the visible and infrared modalities, without considering the reliability of each modality in different tracking scenarios. Therefore, complementarity between two modalities is not fully mined. To address this dilemma, we propose to learn the reliability of each modality using a unified loss function:

$$L(\alpha) = \frac{1}{2} \left\| y - \alpha_v \sum_{c=1}^C \beta_v^c P^T x_v^c * w_v^c - \alpha_i \sum_{c=1}^C \beta_i^c P^T x_i^c * w_i^c \right\|_2^2 + \frac{\sigma}{2} \left\| \alpha - \alpha^r \right\|_2^2\tag{14}$$

where  $\alpha = [\alpha_v, \alpha_i]$  represents the reliability vector and  $\alpha^r$  is the reference vector.  $\sigma$  is the regularization parameter, which can control the change degree of the vector  $\alpha$ . The closed-form solution in the Fourier domain of Eq. (14) is expressed as follows:

$$\alpha = \frac{\hat{M}_\alpha^T \hat{y} + \mu N \alpha^r}{\hat{M}_\alpha^T \hat{M} + \mu N}\tag{15}$$

where  $\hat{M}_\alpha = [\hat{M}_v, \hat{M}_i]$  and  $M_n = \sum_{c=1}^C \beta_n^c P^T x_n^c * w_n^c, n \in [v, i]$ .

## 2.4 Object Localization

The final response map used for localization is acquired by weighted sum the response map of each modality, which can be expressed as follows:

$$\hat{R} = \alpha_v \sum_{c=1}^C \beta_v^c \hat{x}_v^c * \hat{u}_v^c + \alpha_i \sum_{c=1}^C \beta_i^c \hat{x}_i^c * \hat{u}_i^c \quad (16)$$

where  $v$  and  $i$  represent the visible modality and infrared modality, respectively.

## 3 Experiments

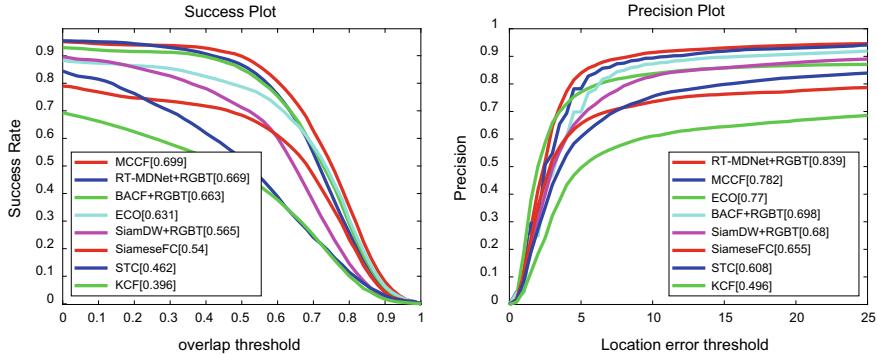
### 3.1 Setups

To verify the effectiveness of the proposed tracker, we perform extensive experiments on GTOT [6] benchmarks. The GTOT benchmarks contains 50 pairs of visible and infrared images aligned in space and time, including 7 attributes. We only use the 31-channel Hog features for feature representation. The learning rate is set to 0.013. The channel and modality regularization parameters are empirically chosen to 0.05 and 0.09, respectively. The reference weights  $\boldsymbol{\alpha}^r$  and  $\boldsymbol{\beta}^r$  are initialized as  $\boldsymbol{\alpha}^r = [1, 1][1, 1]$  and  $\boldsymbol{\beta}^r = [1, \dots, 1]$ , respectively. Experiments are performed on a PC, equipped with an Intel i7-8700 K CPU (3.7 GHz) and a single RTX2080Ti GPU. The area under the curve (AUC) and distance precision (DP) in one-pass evaluation (OPE) [7] are adopted for ranking all trackers.

### 3.2 Quantitative Evaluation

We compare the proposed tracker with 7 state-of-the-art trackers, including RGB trackers, i.e., SiameseFC [8], ECO [9], KCF [10], and RGB-T fusion trackers, i.e., BACF + RGBT [11], SiamDW + RGBT [12], STC [13], RT-MDNet + RGBT [14].

**Overall Performance.** Figure 2 shows both success and plots of all trackers. Overall, the proposed tracker achieves almost the best performance. In terms of AUC, our tracker occupies the best result with a score of 0.699. Although the RT-MDNet +



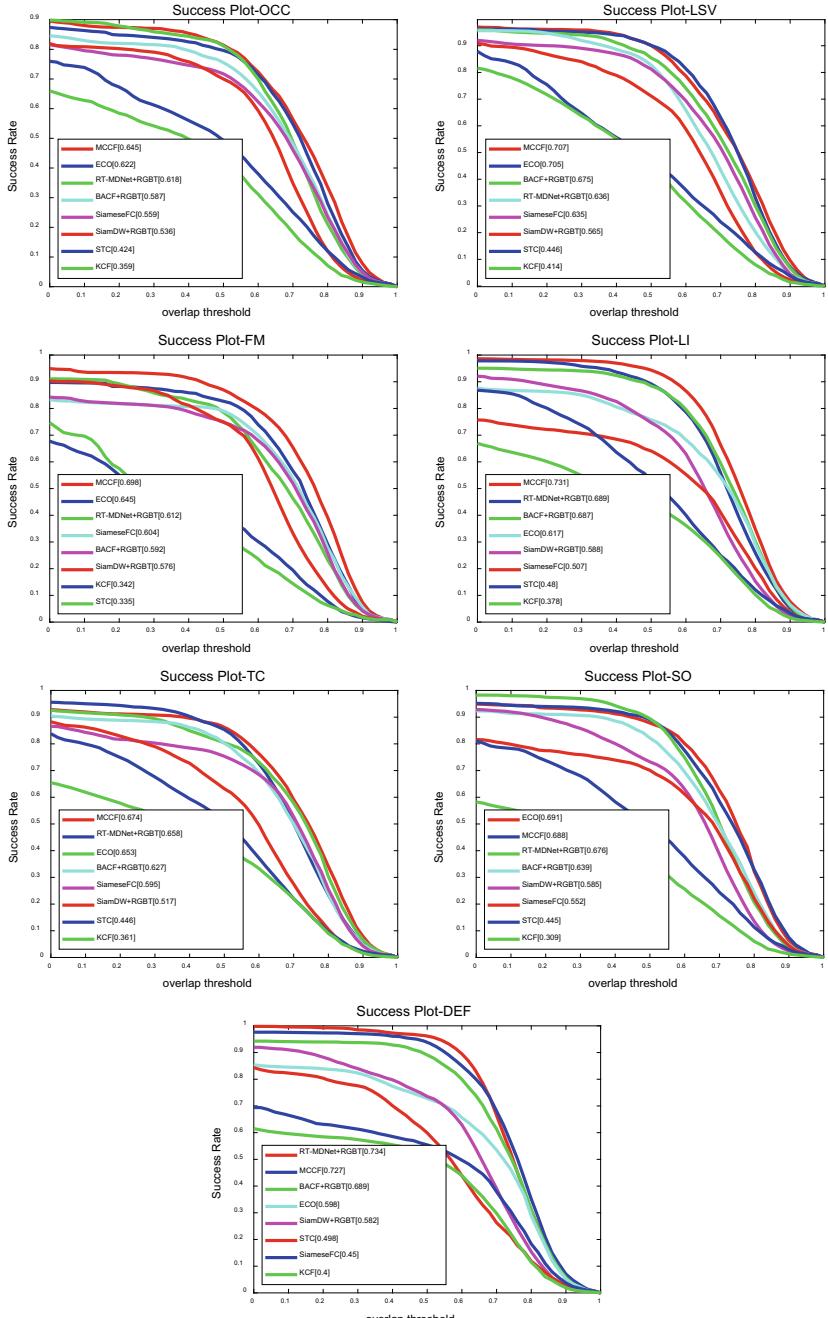
**Fig. 2** Success and precision plots of the proposed tracker and other state-of-the-art trackers on GTOT benchmark. AUC and DP scores of each tracker are reported in the legend

RGBT [14] tracker has the best DP score, the tracking speed is slower than our tracker (26 FPS vs 10FPS).

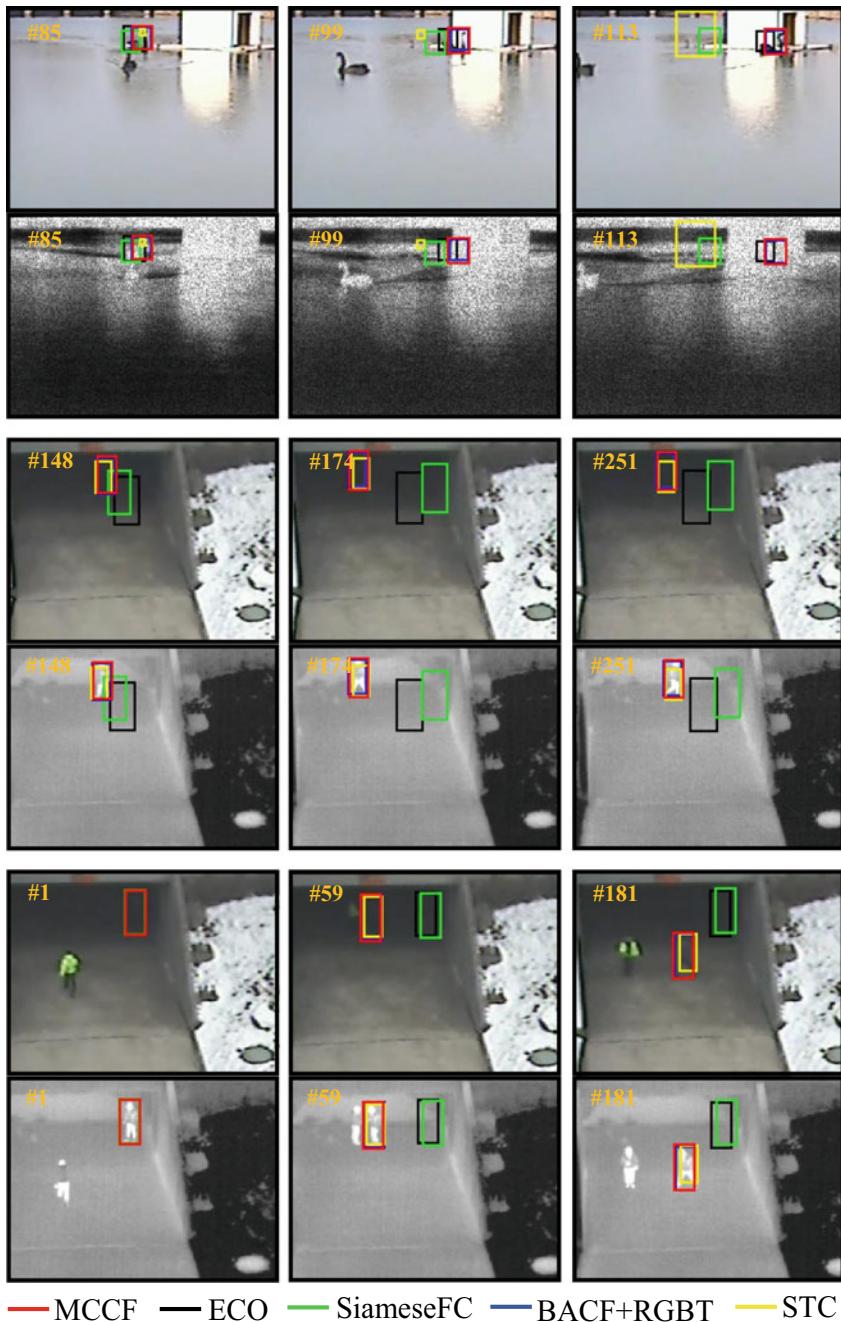
**Attribute Analysis.** The GTOT contains 7 attributes, i.e., occlusion (OCC), large scale variation (LSV), fast motion (FM), low illumination (LI), thermal crossover (TC), small object (SO), deformation (DEF). Figure 3 provides success plots of the proposed tracker and other trackers in various tracking scenarios. Our tracker MCCF outperforms other method in most tracking challenges, especially in fast motion, low illumination and thermal crossover. The results are attributed to the evaluation of the reliability of both visible and infrared modalities in different tracking scenarios and the full use of feature in the channel dimension.

### 3.3 Qualitative Evaluation

We select three representative RGB-T videos from GTOT benchmark to provide visualized comparison, as shown in Fig. 4. In sequence BlackSwan1, the visible modality is more reliable than the infrared modality as the infrared image has the attribute of thermal crossover. Contrary to sequence BlackSwan1, the tracking task is easy to success using the infrared modality the infrared image is not sensitive to light illumination. In sequence Tunnel, the visible and infrared modalities have good complementarity with each other. The visible modality has the color characteristics of both the target and similar target while the infrared modality can capture the target in poor illumination. It can be proved that MCCF can achieve robust tracking in different reliability conditions, which is attributed to the adaptive fusion strategy and the channel reliability.



**Fig. 3** Attribute evaluation on GTOT benchmark. Clearly, AUC score of the proposed method ranks first in the most attributes



**Fig. 4** Visualization of tracking results of the proposed MCCF and 4 other state-of-the-art trackers. From top to bottom: BlackSwan1, GarageHover, Tunnel from the GTOT benchmark, respectively

**Table 1** Ablation analysis on GTOT benchmark

	AUC	DP
Baseline	0.673	0.698
Baseline + CRL	0.685	0.756
Baseline + AMF	0.690	0.768
Ours	0.699	0.782

### 3.4 Ablation Study

To demonstrate the effectiveness of each component, i.e., channel reliability learning (CRL) and adaptive modality fusion (AMF), we develop four RGB-T fusion trackers, *i.e.*, (1) ‘Baseline’ denotes the BACF tracker using response-level fusion; (2) ‘Baseline + CRL’ represents ‘Baseline’ tracker equipped with channel reliability learning; (3) ‘Baseline + AMF’ stands for ‘Baseline’ with adding adaptive modality fusion; (4) ‘Ours’ is the final trackers that combines the ‘Baseline’ with both channel reliability learning and adaptive modality fusion. From Table 1, we can see both CRL and AMF modules can effectively improve tracking performance. Besides, compared to ‘Baseline’, our tracker improves AUC and DP scores by 2.6% and 8.4%, respectively.

## 4 Conclusion

We propose a novel correlation filter framework based on decision-level fusion for RGB-T tracking, which can adjust reliability weights of both visible and infrared modalities and unlock the potential power of inter-channel. Specifically, we propose a single objective function to alternately optimize the filter and channel reliability weight. Furthermore, a novel loss is designed for reliability evaluation of each modality. With the reliability evaluation of both channel and modality dimension, the proposed tracker has achieved promising performance on the GTOT benchmark.

## References

1. Stephen, R.S., Alex, L.C.: Enhanced target tracking through infrared-visible image fusion. In: 14th International Conference on Information Fusion, pp. 1–8. IEEE, Chicago, IL, USA (2011)
2. Alex, L.C., Stephen, R.S.: Fusing concurrent visible and infrared videos for improved tracking performance. Opt. Eng. **52**(1), 7004 (2013)
3. Lichao, Z., Martin, D., Abel, G.G., Joost, W., Fahad, S.H.: Multi-modal fusion for end-to-end RGB-T tracking. In: ICCV Workshop, pp. 2252–2261. Springer, Seoul, Korea (2019)
4. Chenglong, L., Chengli, Z., Yan, H., Jin, T., Liang, W.: Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In: Proceedings of European Conference on Computer Vision, pp. 831–847. Springer, Munich, Germany (2018)
5. Yulong, W., Chenglong, L., Jin, T.: Learning soft-consistent correlation filters for RGB-T object tracking. In: PRCV 2018, pp. 259–306. Springer, Guangzhou, China (2018)

6. Chenglong, L., Hui, C., Shiyi, H., Xiaobai, L., Jin, T., Liang, L.: Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE TIP* **25**(12), 5743–5756 (2016)
7. Yi, W., Jongwoo, L., Ming-Hsuan, Y.: Online object tracking: a benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2411–2418. Portland Oregon (2013)
8. Luca, B., Jack, V., João, F.H., Andrea, V., Philip, H.S.T.: Fully-convolutional Siamese networks for object tracking. In: ECCV Workshops, pp. 850–865. Springer, Amsterdam (2016)
9. Martin, D., Goutam, B., Fahad, S.K., Michael, F.: ECO: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6638–6646. Springer, Honolulu, Hawaii (2017)
10. João, F.H., Rui, C., Martins, P., Jorge B.: High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 583–596 (2014)
11. Hamed, K. G., Ashton, F., Simon L.: Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1135–1143. Venice, Italy (2017)
12. Zhipeng, Z., Houwen, P.: Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4591–4600. Long Beach, California (2019)
13. Kaihua, Z., Lei, Z., Qingshan, L., David, Z., Ming-Hsuan, Y.: Fast visual tracking via dense spatio-temporal context learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuyte- laars, T. (eds.) *ECCV 2014. LNCS*, vol. 8693, pp. 127–141. Springer, Cham (2014)
14. Ilchae, J., Jeany, S., Mooyeon, B., Bohyung, H.: Real-time MDNet. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision—ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, vol. 11208. Springer, Cham (2018)

# Local Binary Complement Pattern for Color-Inversion Invariant Texture Classification



Yuqian Wu  and Tiecheng Song 

**Abstract** Existing Local Binary Pattern (LBP) methods cannot well handle the color-inversion changes. To address this problem, we propose a novel color descriptor called Local Binary Complement Pattern (LBCP) for color-inversion invariant texture classification. The proposed LBCP consists of three local operators, i.e., LBCP\_S, LBCP\_C and LBCP\_O, all of which use the complement information to achieve the color-inversion invariance. Specifically, LBCP\_S encodes the sign information of local neighboring differences in each channel. LBCP\_C and LBCP\_O encode the binary color values and the ordering information of central pixels, respectively, across color channels. After encoding the image using these three operators, we obtain multiple histograms and concatenate them as the LBCP descriptor. Experiments on several color texture databases demonstrate the effectiveness of LBCP for texture classification under color-inversion changes.

**Keywords** LBP · Color · Features · Texture · Classification

## 1 Introduction

Texture is an important visual feature which is widely studied in the fields of image processing and computer vision. The extraction of texture features has been a hot topic in texture classification, scene recognition and so on [1–4]. For texture classification, it is a challenging task to extract texture features which are discriminative to distinguish images of distinct classes and meanwhile invariant to various image variations [5, 6] in rotation, illumination, scaling, etc.

Local Binary Pattern (LBP), originally proposed by Ojala et al. [1], is one of the well-known texture descriptors. LBP compares each pixel with its neighbors and encodes the resulting binary strings into integers to build a histogram as the image descriptor. Because LBP is robust to illumination changes and has low computational complexity, a large number of LBP variants have been developed in the past few

---

Y. Wu · T. Song ()

Chongqing University of Posts and Telecommunications, Chongqing 400065, China

e-mail: [songtc@cqupt.edu.cn](mailto:songtc@cqupt.edu.cn)

years. For example, Complete LBP (CLBP) [7] jointly encodes three complementary components to improve the classification performance. Local Ternary Pattern (LTP) [8] extends LBP to three-valued quantization to improve the noise robustness. Non-Redundant LBP (NRLBP) [9] takes the minimum of one LBP code and its complement to achieve invariance to foreground and background changes. In the literature, there are some LBP variants developed to handle color images. For example, multichannel adder-based and decoder-based LBPs (maLBP and mdLBP) [10] extract cross-channel features to describe color images. In Ref. [11], LBP for color images (LBPC) uses a plane in 3D color space to threshold color pixels into two categories and encodes the resulting binary patterns to form histogram features. In Ref. [12], Spatially Weighted Order Binary Pattern (SWOBP) combines color orders and local color differences to encode cross-channel color pixels. In Ref. [13], Quaternionic Local Ranking Binary Pattern (QLRBP) extracts cross-channel color features in the quaternionic domain.

Despite the above progress, none of the above descriptors addresses the color-inversion problem for color image description. As shown in Fig. 1, the extraction of color-inversion invariant features is crucial to successfully classify, retrieve and match these color images. Note that NRLBP was designed to address the grayscale inversion of gray images, not for color images. If we directly concatenate the NRLBP features from each color channel, the classification performance, as shown in our experiments, is unsatisfactory because the color correlation information is not sufficiently captured.

In this paper, we propose a novel Local Binary Complement Pattern (LBCP) descriptor to address the color-inversion problem. Our idea is to encode local spatial information and cross-channel color features and make use of the complement codes to achieve color-inversion invariance. Specifically, we propose three local operators, i.e., LBCP\_S (Signs of local differences), LBCP\_C (Central pixels), and LBCP\_O (color Orderings among channels), which form multiple histograms to



**Fig. 1** Top: original images. Bottom: the corresponding images with color-inversion changes

capture discriminative color features. Experiments for color texture classification show that LBCP significantly outperforms other descriptors under color-inversion changes.

## 2 Review of LBP

The basic LBP operator [1] for a given central pixel is defined as follows

$$LBP_{r,P} = \sum_{p=0}^{P-1} s(g_{r,p} - g_c) 2^p \quad (1)$$

where  $g_c$  and  $g_{r,p}$  ( $p = 0, 1, \dots, P-1$ ) are the gray values of the central pixel and the  $p$ -th neighboring pixel, respectively (there are  $P$  pixels evenly distributed on a circle of radius  $r$ ). The sign function  $s(x)$  is defined as

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2)$$

Other LBP encoding schemes include uniform LBP ( $LBP_{r,P}^{u2}$ ) and rotation invariant uniform LBP ( $LBP_{r,P}^{riu2}$ ) [1]. For  $LBP_{r,P}^{u2}$ , a uniformity measure  $U$  is defined as the number of spatial transitions (0/1 changes) in a binary string, i.e.,

$$\begin{aligned} U(LBP_{r,P}) &= |s(g_{r,P-1} - g_c) - s(g_{r,0} - g_c)| \\ &\quad + \sum_{p=1}^{P-1} |s(g_{r,p} - g_c) - s(g_{r,p-1} - g_c)| \end{aligned} \quad (3)$$

In Ref. [1], Ojala et al. defined patterns with a  $U$  value of less than or equal to 2 as uniform patterns, and others as non-uniform patterns. Hence, there are  $P(P-1) + 3 LBP_{r,P}^{u2}$  codes.

For rotation invariant image description, the  $LBP_{r,P}^{riu2}$  operator is defined as:

$$LBP_{r,P}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_{r,p} - g_c), & U(LBP_{r,P}) \leq 2 \\ P + 1, & \text{otherwise} \end{cases} \quad (4)$$

Hence, there are  $P + 2 LBP_{r,P}^{riu2}$  codes in total.

Generally speaking, under linear color-inversion changes, the LBP codes defined in (1) for each color channel will be changed to their complements (i.e., the bits 1 and 0 in a binary string will be exchanged). Based on this, NRLBP [9] takes the minimum of each LBP code and its complement to obtain invariance. In this paper, we explore

the complement operation to hand color images and construct a discriminative color descriptor LBCP while achieving the robustness to color-inversion changes. To our knowledge, we are the first to extract color-inversion invariant image features.

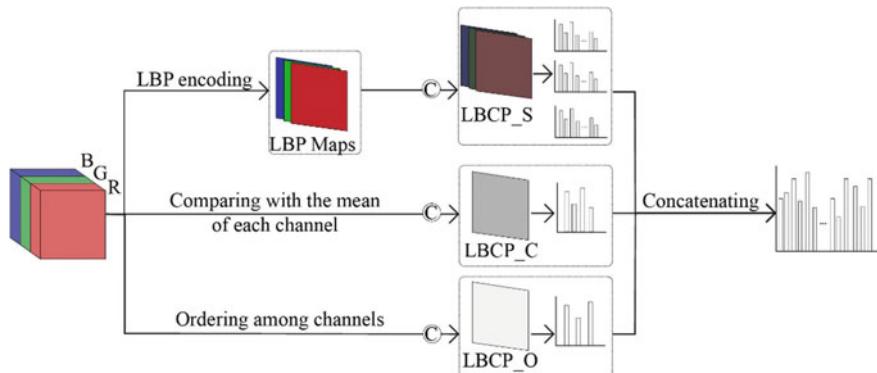
### 3 Proposed Color Descriptor

Figure 2 shows the flowchart of constructing the LBCP descriptor from an RGB color image. LBCP contains three local operators: LBCP\_S (Signs of local differences for each color channel), LBCP\_C (Central pixels across color channels), and LBCP\_O (color Orderings across color channels). Among them, LBCP\_S forms three code maps and each of the other two operators forms one code map. Thus, we obtain five histograms which are concatenated as the final LBCP descriptor.

#### 3.1 LBCP\_S Operator

Under linear color-inversion changes, the LBP codes defined in (1) for each color channel will be changed to their complements. Following NRLBP [9], we compute LBP codes on each channel separately and compare their complements to achieve color-inversion invariance. The resulting histograms are low-dimensional and capture local spatial relationships. Note that when the color inversion occurs, non-uniform patterns remain non-uniform (for both “ $u2$ ” and “ $riu2$ ”). Hence, we encode all non-uniform patterns into a unique code.

Formally, the code of  $LBCP_{S_{r,p}}^{u2}$  ( $X$  refers to R, G or B) is defined as



**Fig. 2** Flowchart of constructing the LBCP descriptor (© denotes the complement operation)

$$LBCP\_S_{r,P}^{u2}\_X = \begin{cases} \min\{LBP_{r,P}^{u2}, P(P-1)+1-LBP_{r,P}^{u2}\}, & U(LBP_{r,P}) \leq 2 \\ (P(P-1)+2)/2, & otherwise \end{cases} \quad (5)$$

The code of  $LBCP\_S_{r,P}^{riu2}\_X$  is defined as

$$LBCP\_S_{r,P}^{riu2}\_X = \begin{cases} \min\{LBP_{r,P}^{riu2}, P - LBP_{r,P}^{riu2}\}, & U(LBP_{r,P}) \leq 2 \\ P/2 + 1, & otherwise \end{cases} \quad (6)$$

Based on Eqs. (5) [or (6)], we can build a histogram of  $LBCP\_S_{r,P}^{u2}\_X$  (or  $LBCP\_S_{r,P}^{riu2}\_X$ ) codes and then concatenate all histograms from RGB channels as the LBCP\_S descriptor. For (5) and (6), the dimensions of corresponding descriptors are  $((P(P-1)+2)/2+1) \times 3$  and  $(P/2+2) \times 3$ , respectively.

### 3.2 LBCP\_C Operator

The central pixels contain useful information for image description. In CLBP [7], each central pixel is binarized to capture the local and global contrast relationships of the image. Mathematically, CLBP\_C (Central pixels) is defined as

$$CLBP\_C = s(g_c, c_I) \quad (7)$$

where  $c_I$  is the mean gray value of all pixels in image  $I$ , and  $s(x)$  is defined in (2).

For the proposed LBCP\_C operator, we extend CLBP\_C to deal with color images. Specifically, we jointly encode the values of CLBP\_C for all RGB channels (a total of  $2^3 = 8$  possible values) and based on the corresponding complement code compute the final encoding value. Given a central pixel, the LBCP\_C code is obtained through

$$LBCP\_C = \min(C_{joint}, 7 - C_{joint}) \quad (8)$$

where  $C_{joint} = 0, \dots, 7$  denotes the joint encoding of CLBP\_C for all RGB channels, i.e.,

$$C_{joint} = CLBP\_C\_R + CLBP\_C\_G \times 2 + CLBP\_C\_B \times 4 \quad (9)$$

Accordingly, we can construct a LBCP\_C histogram whose dimension is 4.

**Table 1** The LBCP\_O codes of central pixels in RGB images

$O(g_C)$	Code	$O(g_C)$	Code
1, 2, 3	0	3, 2, 1	0
1, 3, 2	1	3, 1, 2	1
2, 1, 3	2	2, 3, 1	2

### 3.3 LBCP\_O Operator

The above two operators are based on binary comparisons (reflecting partial ordering relationships) which cannot capture the full ordering relationships [12, 14]. To capture such relationships among the RGB channels and meanwhile preserve invariance to color inversion, we propose an additional LBCP\_O (color Ordering) operator.

Given a central pixel, the RGB values are given as

$$g_C = (g_R, g_G, g_B) \quad (10)$$

We assign an encoding value to this central pixel according to Table 1. In this table,  $O(g_C)$  describes the ordering of the elements in  $g_C$  in ascending order. For example, for  $g_C = (30, 70, 10)$ , we have  $O(g_C) = (2, 3, 1)$  and the corresponding LBCP\_O code is 2. When an RGB image is reversed, we need to encode the complement of each order to the same code. For example, for  $g_C' = (-30, -70, -10)$ , we have  $O(g_C') = (2, 1, 3)$  and the corresponding LBCP\_O code should also be 2. In this way, we can achieve the color-inversion invariance.

Accordingly, we can construct a LBCP\_O histogram whose dimension is 3.

### 3.4 LBCP Descriptor

To construct a compact LBCP descriptor, we first obtain three LBCP\_S code maps (i.e., LBCP\_S\_R, LBCP\_S\_G, and LBCP\_S\_B) and two LBCP\_C and LBCP\_O code maps (each with one code map). Then, we build five histograms based on the code maps and concatenate all histograms as the LBCP descriptor (see Fig. 2). For  $riu2$  patterns, the descriptor dimension is  $((P(P - 1) + 2)/2 + 1) \times 3 + 7$ . For  $u2$  patterns, the descriptor dimension is  $(P/2 + 2) \times 3 + 7$ . In the proposed LBCP descriptor, LBCP\_S encodes the local spatial relationships for each color channel while LBCP\_C and LBCP\_O encode cross-channel color features. The resulting LBCP descriptor is not only robust to color-inversion changes but also discriminative for color image description.

## 4 Experiments

To evaluate the proposed color descriptor, we use three well-known color texture databases, i.e., Outex-TC30, Outex-TC31, and Outex-TC10-c. We normalize the color values in each color channel (R, G and B) into [0, 255] and use the nearest neighbor classifier with the chi-square distance for color texture classification. We evaluate different descriptors in terms of the classification accuracy.

### 4.1 Databases and Experimental Setup

**Outex-30:** It contains 68 color textures, each with 20 images. The rotation angles of images are  $0^\circ$ ,  $5^\circ$ ,  $10^\circ$ ,  $15^\circ$ ,  $30^\circ$ ,  $45^\circ$ ,  $60^\circ$ ,  $75^\circ$ , and  $90^\circ$ . The illumination type is ‘Inca’ and the resolution is 100 dpi.

**Outex-10-c:** It contains 24 color textures, each with 20 images and the rotation angles are  $0^\circ$ ,  $5^\circ$ ,  $10^\circ$ ,  $15^\circ$ ,  $30^\circ$ ,  $45^\circ$ ,  $60^\circ$ ,  $75^\circ$ , and  $90^\circ$ . The illumination type is ‘Inca’ and the resolution is 100 dpi.

**Outex-31:** It contains 68 color textures, each with 20 samples. The illumination type is ‘Inca’ and the resolution is 100 and 120 dpi.

The above three databases have the predefined training and testing sets and the first two ones are used to evaluate rotation invariance (<http://lagis-vi.univ-lille1.fr/datasets/outex.html>). To model color-inversion changes, we transform the color values of all samples in the testing set of the three databases by

$$I' = -\lambda \times I + 255 \quad (11)$$

where  $\lambda$  is a randomly number whose value is in the range of 0 and 1.

We implement LBP [1], LTP [8], CLBP [7], NRLBP [9] and our LBCP using *riu2* encoding for Outex-TC30 and Outex-TC10-c while we implement them using *u2* for Outex-TC31. For the first four methods, we obtain the final image features by concatenating the histograms extracted from each color channel. For SWOBP [12], LBPC [11], QLRBP [13], maLBP [10] and mdLBP [10], we directly use the default configurations to handle color images as suggested in the original papers.

### 4.2 Classification Results and Analysis

Tables 2 and 3 list the classification accuracies of different descriptors on Outex-TC30 and Outex-TC10-c with rotation changes. Firstly, the classification accuracies of LBP, CLBP and LTP are pretty good on the original databases. However, their performance

**Table 2** Classification accuracies (%) on the Outex-TC30 database

Descriptors	(R, P)	Original	Color-inversion changes
LBP [1]	(2, 16)	86.35	36.34
CLBP [7]	(2, 16)	86.21	33.06
LTP [8]	(2, 16)	93.18	38.44
SWOBP [12]	–	62.73	22.71
LBPC [11]	–	43.41	18.14
QLRBP [13]	–	56.34	27.74
maLBP [10]	(2, 16)	51.82	25.05
mdLBP [10]	(2, 16)	56.57	33.11
NRLBP [9]	(2, 16)	82.29	80.03
LBCP	(2, 16)	86.63	<b>85.44</b>

Bold values indicates the best result obtained by our method

**Table 3** Classification accuracies (%) on the Outex-TC10-c database

Descriptors	(R, P)	Original	Color-inversion changes
LBP[1]	(2, 16)	96.32	34.92
CLBP[7]	(2, 16)	97.29	40.44
LTP[8]	(2, 16)	98.90	37.18
SWOBP[12]	–	74.01	21.66
LBPC[11]	–	50.15	27.42
QLRBP[13]	–	64.27	38.61
maLBP[10]	(2, 16)	60.54	36.27
mdLBP[10]	(2, 16)	66.19	49.03
NRLBP[9]	(2, 16)	93.95	92.86
LBCP	(2, 16)	97.47	<b>97.21</b>

Bold values indicates the best result obtained by our method

drops dramatically under color-inversion changes. This is because they use signed neighboring differences which are sensitive to color-inversion changes. Secondly, the proposed LBCP achieves the best performance on these two databases under color-inversion changes, demonstrating its discriminative ability for texture description and its robustness to color inversion and image rotation. Thirdly, although NRLBP is designed to solve the color-inversion problem (also using the idea of complement), its classification accuracies are inferior to ours in these two rotated Outex databases. This indicates that the concatenated histogram features extracted from three separate RGB channels are not very discriminative for color texture description.

Table 4 shows the classification accuracies of different descriptors on the Outex-TC31 database without rotation changes. Firstly, SWOBP, mdLBP and maLBP are the top three methods on the original Outex-TC31 database, achieving the classification accuracies of 93.23%, 93.01% and 91.17%, respectively. These results

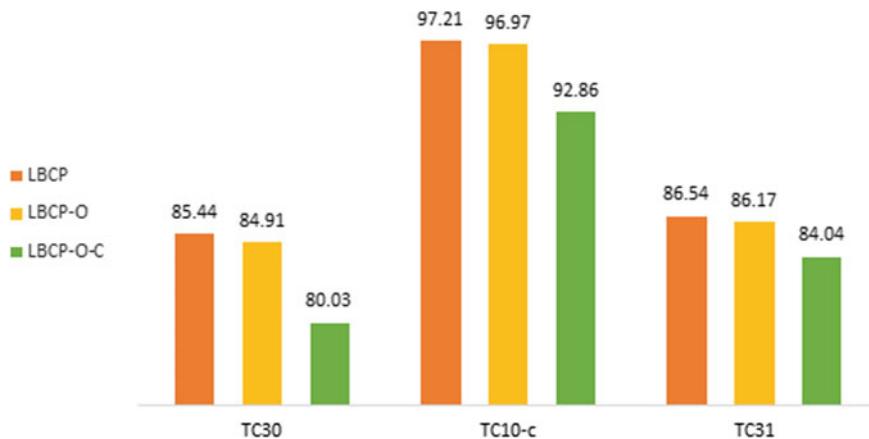
**Table 4** Classification accuracies (%) on the Outex-TC31 database

Descriptors	(R, P)	Original	Color-inversion changes
LBP [1]	(2, 16)	89.55	42.35
CLBP [7]	(2, 16)	85.73	40.29
LTP [8]	(2, 16)	90.22	43.08
SWOBP [12]	–	93.23	32.86
LBPC [11]	–	84.41	31.10
QLRBP [13]	–	81.47	39.41
maLBP [10]	(2, 16)	91.17	39.19
mdLBP [10]	(2, 16)	93.01	48.67
NRLBP [9]	(2, 16)	81.10	84.04
<b>LBCP</b>	(2, 16)	<b>87.13</b>	<b>86.54</b>

Bold values indicates the best result obtained by our method

are much better than those obtained on Outex-TC30 and Outex-TC10-c because SWOBP, mdLBP and maLBP are implemented without rotation invariance. However, their performance becomes very poor under color-inversion changes. Secondly, our LBCP outperforms all other descriptors by a large margin under color-inversion changes. Therefore, LBCP is effective for color-inversion invariant texture description and classification. Thirdly, NRLBP shows the second-best performance under color-inversion changes due to the lack of rich yet discriminative features, as opposed to our LBCP.

Finally, we illustrate the impacts of the three operators in LBCP on the classification accuracy. Figure 3 shows the results where LBCP-O means the LBCP descriptor without the LBCP\_O histogram and LBCP-O-C means the LBCP descriptor without the LBCP\_O and LBCP\_C histograms (i.e. it is reduced to NRLBP). As can be seen,



**Fig. 3** Impacts of the three operators on the classification accuracy (%)

each operator has some important impacts on the classification performance and the combination of all the three operators leads to the best classification performance. In particular, the LBCP\_C operator contributes more to the performance gain. The main reason is that LBCP\_C describes the relationships between each central pixel and the whole image while LBCP\_O only describes the relationships among three central pixels. In this sense, the jointly encoded LBCP\_C features are more discriminative than the LBCP\_O features.

## 5 Conclusions

This paper presents LBCP for color-inversion invariant texture classification. With the use of the complement information, three local operators (i.e., LBCP\_S, LBCP\_C and LBCP\_O) are developed to encode color features and achieve color-inversion invariance. LBCP\_S encodes the sign information of local differences in each color channel. LBCP\_C and LBCP\_O encode the binary color values and the ordering information of central pixels, respectively, across color channels. The compact LBCP descriptor is formed by concatenating histograms obtained from all the code maps. Experiments for color texture classification demonstrate the superiority of LBCP under color-inversion changes.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (61702065) and by the Chongqing Research Program of Basic Research and Frontier Technology (cstc2018jcyjAX0033). We thank the reviewers for improving the quality of this paper.

## References

1. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
2. Tiecheng, S., Hongliang, L., Fanmang, M., Qingbo, W., Bing, L.: Exploring space-frequency co-occurrences via local quantized patterns for texture representation. *Pattern Recogn.* **48**(8), 2621–2632 (2015)
3. Xiuwen, L., Deliang, W.: Image and texture segmentation using local spectral histograms. *IEEE Trans. Image Process.* **15**(10), 3066–3077 (2006)
4. Tiecheng, S., Hongliang, L.: WaveLBP based hierarchical features for image classification. *Pattern Recogn. Lett.* **34**(12), 1323–1328 (2013)
5. Kandaswamy, U., Schuckers, S.A., Adjerooh, D.: Comparison of texture analysis schemes under nonideal conditions. *IEEE Trans. Image Process.* **20**(8), 2260–2275 (2011)
6. Tiecheng, S., Hongliang, L., Fanmang, M., Qingbo, W., Jianfei, C.: LETRIST: Locally encoded transform feature histogram for rotation-invariant texture classification. *IEEE Trans. Circuits Syst. Video Technol.* **28**(7), 1565–1579 (2018)
7. Zhenhua, G., Lei, Z., David, Z.: A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.* **19**(6), 1657–1663 (2010)

8. Xiaoyang, T., Bill, T.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **19**(6), 635–1650 (2010)
9. Duc, T.N., Zhimin, Z., Philip, O., Wanqing, L.: Object detection using non-redundant local binary patterns. In: 17th IEEE International Conference on Image Processing, pp. 4609–4612. IEEE, Piscataway, NJ (2010)
10. Dubey, S.R., Singh, S.K., Singh, R.K.: Multichannel decoded local binary patterns for content-based image retrieval. *IEEE Trans. Image Process.* **25**(9), 4018–4032 (2016)
11. Singh, C., Walia, E., Kaur, K.P.: Color texture description with novel local binary patterns for effective image retrieval. *Pattern Recogn.* **76**, 50–68 (2018)
12. Tiecheng, S., Jie, F., Shiyan, W., Yurui, X.: Spatially weighted order binary pattern for color texture classification. *Expert Syst. Appl.* **147**, 113167 (2020)
13. Rushi, L., Yicong, Z., Yuanyan, T.: Quaternionic local ranking binary pattern: a local descriptor of color images. *IEEE Trans. Image Process.* **25**(2), 566–579 (2016)
14. Tiecheng, S., Jie, F., Lin, L., Chenqiang, G., Hongliang, L.: Robust texture description using local grouped order pattern and non-local binary pattern. *IEEE Trans. Circuits Syst. Video Technol.* **31**(1), 189–202 (2021)

# Video Instance Segmentation of Rock Particle Based on MaskTrack R-CNN



Man Chen , Maojun Li , and Yiwei Li

**Abstract** Video instance segmentation (VIS) of rock particles in motion is the basis for revealing the laws of motion and quantitative analysis. It has important scientific and engineering value. Use an end-to-end network called MaskTrack R-CNN to complete the VIS task for rock particles. The network introduces a new tracking branch on Mask R-CNN. It integrates particle detection, segmentation, and tracking tasks into the framework. The tracking branch primarily uses appearance similarity cues to linearly combine cues such as semantic consistency and spatial correlation to improve tracking accuracy. To facilitate the study of rock particle visibility, we have created a set of experimental equipment for collecting rock particle datasets. We conducted training and testing experiments to verify the effectiveness of the algorithm and compared it to some baselines of our own dataset. Experimental results show that MaskTrack R-CNN uses ResNet-50 to get 33.1% AP. It better than other two-stage models. This work provides an intelligent solution for meso-analyzing particles.

**Keywords** Video instance segmentation · Rock particle · MaskTrack R-CNN

## 1 Introduction

Video instance segmentation is a challenging visual task. You need to track the instances across frames and segment objects in individual frames. Many video-based tasks have core applications, such as video editing, autonomous driving, and augmented reality. The pioneering work of VIS is MaskTrack R-CNN [1], which is an extension of Mask R-CNN [2]. In addition to the initial three branches of object classification, bounding box regression, and masking, there is a fourth branch with external storage for tracking object instances frame by frame. First, the use of Region Proposal Network (RPN) [3] of Faster R-CNN to generate a set of candidate recommendations. Then, motion-based ROI features are clipped and inserted at the beginning of each task for bounding box prediction, object masking, and object

---

M. Chen · M. Li · Y. Li

Changsha University of Science and Technology, Changsha 410114, China

e-mail: [19205060770@stu.csust.edu.cn](mailto:19205060770@stu.csust.edu.cn)

tracking. It also recommends a large video dataset called YouTube-VIS to measure video version segmentation algorithms. The new dataset can be used as a useful benchmark for various pixel-level video comprehension tasks.

Particles are an important part of global geographic disasters and are used in construction projects and vehicles [4, 5]. The development of general theory of various materials was one of the 125 cutting-edge science projects. Interpreting the rock particles in motion is the basis for demonstrating the laws of motion and their size parameters, and can provide accurate guidance for construction work. This is also the reason for the insufficient use of verification studies and numerical modeling methods (such as Finite Element Method and Discrete Element Method) in engineering technology. When you generalize these models [6, 7], similar particles can also provide reliable data for the VIS particle mode. In short, the clarity of rock light research is of great value to scientific and technological research.

In this article, we will introduce VIS into digital technology and apply the rock particle video segmentation method on MaskTrack R-CNN. In summary, the main contributions of this work are as follows:

- An end-to-end approach is used to achieve particle visibility by integrating detection, segmentation and tracking operations into the video frames. As far as we know, this is the first VIS application in construction.
- We develop experimental procedures for capturing video and create a dataset with moving objects under external force, which includes 160 videos of rock particles.
- We conducted training and measurement experiments to determine the efficiency of the process and compared it with the different bases of our self-designed data set.

All our papers were designed this way. In Sect. 2 we briefly describe the work of VIS and the main parts that are developed. In Sect. 3 we officially discuss MaskTrack R-CNN algorithms. Section 4 introduces a new set of equipment and experimental results.

## 2 Related Work

Little research has been done on VIS, especially on the VIS components. However, particle separation has been thoroughly analyzed as the basis for VIS to create some new machine vision algorithms. In this section, we will look at the development.

### 2.1 Video Instance Segmentation

VIS requires simultaneous classification, distribution and tracking in the video. Depending on how the sequence originates, it can be divided into two types. One type divides tracking and search into two parts [1, 8, 9]. In the form of research,

examples are set up in a frame-by-frame manner using existing image-level example segmentation methods. And the detected positions can be linked to different frames of the tracking component. The second type is abbreviated as ‘Clip-Match’ methods [10, 11]. Breaks the whole video into several short clips and creates a separate VIS for each by scattering or space-time embedding. It then links the clips to other related clips.

## 2.2 *Particle Image Processing*

Particle photo processing involves the classification, detection and segmentation of particles. Extracting particle mask is the basis of VIS, so we summarize it mainly in this single copy. Particle images often have density and adhesion properties which make the distribution process very difficult [12]. It focuses on solving the problem of pollution caused by interactions and shadows in metal images and designs a method of distributing metal images based on holistic nested edge detection [13]. The light-weight U-Net deep training network is designed to automatically detect particles from photographs and obtain potential particle change maps. This method can be used to monitor the particle product quality. Liu et al. [14] introduced a photo sharing method based on U-Net and its improved network. Pre-processed real-time images from open cast mines to reduce noise and capture the object area with applicable traditional vision techniques.

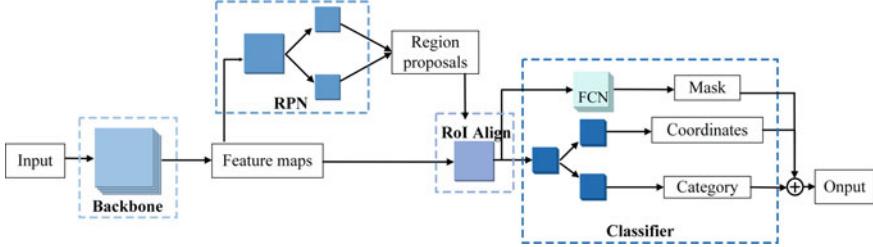
## 3 MaskTrack R-CNN

The Mask R-CNN is the basis of the MaskTrack R-CNN, so we show the Mask R-CNN first and then explain the new parts in detail.

### 3.1 *Mask R-CNN*

Mask R-CNN [2] can distribute pixel level images by combining the advantages of object detection network and semantic distribution network. The RoI Align is used to place the RoI Pooling on the R-CNN, which solves the problem of field misalignment. Backbone, RPN, RoI Align and classification: The general network structure of the Mask R-CNN is shown in Fig. 1, which is actually made up of four components.

The backbone is a series of mixed layers that can map features. It has several constitutional layers. Samples are reliable, consistent, and dynamic at every level and receive maps displayed at different sizes. The feature pyramid network (FPN)



**Fig. 1** Mask R-CNN network structure

[15] can be used to obtain multi-layer semantic fuses and post-fusion mapping of different sizes.

RPN is a network that can issue regional proposals for follow-up tasks. Inputs of RPN are the results of the final FPN list. First, it produces a certain number of anchors per pixel on these maps. The probability that these anchors are placed in front or in the background is calculated and the migration is between the combination of these anchors and the corresponding fact on earth. Redistribution and regression can also be tested in terms of RPN loss function. Finally, you will find the appropriate area recommendations and weight parameters after multiple repetitions.

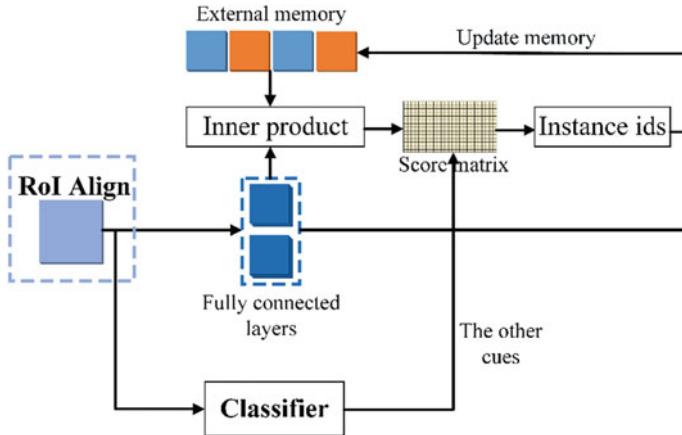
RoI alignment is applied to the instead of approximation in the original RoI section. It can change the location of the search to make a map size without losing location information, so that each pixel remains precisely pointed.

Classifier has three similar branches. Two of these components and layouts can get the more accurate box. One branch uses the Fully Convolution Net (FCN) [16] to predict the mask.

### 3.2 The Tracking Branch

The network adopts a framework in two stages. In the second step, we added a fourth branch to assign a sample label to each candidate box. As shown in Fig. 2, this branch is parallel to the three branches (the above three branches). Let the number of cases recognized by the model from the previous frame is  $n$ . If the new candidate box is one of the previous copies, it will only be entered for  $n$  identities. If this is a new instance, it will be given a new identity. This is generally a distribution problem with several classes. There is a  $n + 1$  class digits number that already identifies the  $n$  instance. A new example is represented by the number 0. The probability of assigning label  $m$  to a candidate box  $i$  is defined as

$$C_i(n) = \begin{cases} \frac{e^{T_{im}}}{1+\sum_{j=1}^n e^{T_{ij}}}, & 1 < m < n \\ \frac{1}{1+\sum_{j=1}^n e^{T_{ij}}}, & m = 0 \end{cases} \quad (1)$$



**Fig. 2** The tracking branch

where  $t_i$  and  $t_j$  are the new features extracted by the tracking branch. Our tracking branch has two layers that are fully connected. These layers can project the feature maps drawn by ROI Align into new features. The two fully connected layers transform the input function cards into 1-D 1024 dimensions. Cross entropy loss is used for a tracking branch, which can be expressed as follows:

$$L_{track} = - \sum_i \log(C_i(r_i)) \quad (2)$$

where  $r_i$  the ground truth instance label.

It also uses external memory for storage for greater efficiency. External memory is dynamically updated as the instance label is assigned to the new candidate frame. If the selected frame belongs to an existing instance, the instance characteristics stored in memory are updated with the new candidate characteristics. If the candidate is given 0 points, the candidate's characteristics are stored in memory and increase by 1 depending on the number of examples identified.

The tracking branch mainly uses the appearance similarity to accomplish the tracking task. However, there are also other information such as semantic consistency, spatial correlation and detection confidence which could be leveraged to determine the instance labels. MaskTrack R-CNN also combines all these cues together to improve the tracking accuracy in a post-processing way. It completes the matching of instances by calculating the score of assigning label  $m$  to the candidate box  $i$  and the calculation formula is shown as follows:

$$L = L_{cls} + L_{box} + L_{mask} + L_{track} \quad (3)$$

### 3.3 The Other Cues

The control branch uses visual equations to perform tasks. However, there are other information that can be used to identify sample tags, such as semantic similarity, positional relationship, and recognition reliability. MaskTrack R-CNN collects all these signals to confirm the accuracy of post-processing. Now complete the random adjustment by calculating the given point of the corner  $m$  of candidate  $i$ , the calculation formula is as follows:

$$S_i(m) = \log C_i(m) + \alpha \log(d_i) + \beta IoU(b_i, b_m) + \gamma \varphi(c_i, c_m) \quad (4)$$

where  $i$  is the sequence number of the candidate box.  $b_i$ ,  $c_i$  and  $d_i$  denote the bounding box prediction, category label and detection score.  $c_m$  is the bounding box prediction and category label associated with the saved features in the memory.  $\varphi(c_i, c_m)$  is a Kronecker delta function which equals 1 when  $c_i$  and  $c_m$  are same and 0 otherwise.  $\alpha$ ,  $\beta$  and  $\gamma$  are hyperparameters which can balance the effect of different cues.

## 4 Experiments

### 4.1 Experimental Equipment and Dataset

We designed a set of test tools to collect data from the parts needed for the experiment. This mainly includes test bench, motion providing device, sight sensor and object to be measured. The motion providing device moves the objects and gives some scrolling action. Test bench is used to secure the testing process. The visual sensor can capture video of the experimental environment. Figure 3 shows a video of the experimental environment captured by the visual sensor.



**Fig. 3** Experimental equipment

The collected videos are used to support our approach, and are divided into different three categories according to the complexity of the data. General categories 9. Each category has 20 different compressed videos. Like the YouTube VIS design guidelines, some objects are defined by manually tracking the boundaries of every 5 frames, and each video has a rate of 30 frames per second. We also change the original frame sizes to  $640 \times 360$  in training. In particular, we divide the data into training and validation functions according to a certain proportion.

## 4.2 Implementation Details

Model training and testing experiments were performed on Ubuntu 18.04. The processor is Intel Core i7-8700 K CPU @ 3.7 GHz and the GPU is NVIDIA GeForce RTX 2080Ti. Use the original MaskTrack R-CNN weights as direct weights to increase the convergence speed. The model is ready by the end of the eighth epoch. The primary learning level is 0.05. The hyperparameters  $\alpha$ ,  $\beta$  and  $\gamma$  are chosen to be 1, 2 and 10 respectively.

## 4.3 Main Results

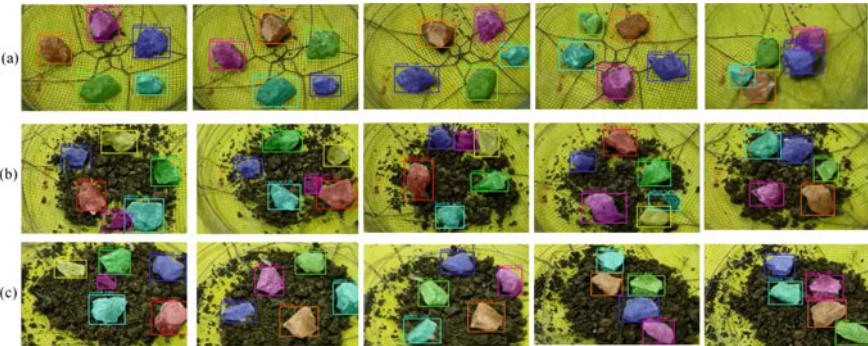
**Baselines** In this experiment, we set up three baselines to compare with MaskTrack R-CNN. They are IoUTrack+ [1], OSMN [17] and DeepSORT [18] respectively. For the frame-by-frame examples generated by the Mask R-CNN, the baseline has the same segmentation effect. Convert the generated video data set into an image. Then create an image data set to train a Mask R-CNN. The structure of the Mask R-CNN is like a network. In addition to the branches of the track.

**Quantitative Results** MaskTrack R-CNN baseline was compared against a self-derived data set. Table 1 shows the results of the MaskTrack R-CNN comparison, achieved across all metrics. Specifically, the MaskTrack R-CNN consistently surpasses baselines by a significant margin in AP (26.4% vs. 33.1% with IoUTrack+,

**Table 1** Quantitative evaluation of the proposed algorithm and baselines

Method	Backbone	AP	AP <sub>50</sub>	AP <sub>75</sub>	AR <sub>1</sub>	AR <sub>10</sub>
IoUTrack+ [1]	ResNet-50	26.4	43.8	27.9	22.3	27.8
OSMN [17]	ResNet-50	29.6	48.0	30.3	21.9	27.2
DeepSORT [18]	ResNet-50	28.7	47.1	32.5	22.4	29.8
MaskTrack R-CNN [1]	ResNet-50	33.1	54.3	36.9	25.4	31.7
MaskTrack R-CNN [1]	ResNet-101	<b>34.8</b>	<b>56.7</b>	<b>38.2</b>	<b>27.5</b>	<b>33.9</b>

The best results are highlighted in bold



**Fig. 4** Sample results of VIS. Each row have six sampled frames from a video sequence

29.6% vs. 33.1% with OSMN and 28.7% versus 33.1% with DeepSORT). In the case of AR, the results of all methods are very low. This may be due to the obvious difference between the background and foreground in the video. These conditions can cause difficulty in segmentation and can also cause particles to disappear. Among them, MaskTrack R-CNN is at least 3.5% higher than its starting lineup. In general, MaskTrack R-CNN can achieve better results than baseline in AR. Moreover, we observe that the ResNet-101 is better than ResNet-50 (33.1% vs. 34.8%).

**Qualitative Results** Figure 4 shows the qualitative results of both videos. Most rock particles are clear, but some are incomplete. Particles, which are mostly blocked, have more serious errors. The reason for this phenomenon is that objects hidden surrounding impurities make classification, segmentation and tracking difficult. In general, MaskTrack R-CNN can segment the most viewed particles.

## 5 Conclusion

In this study we used a holistic method to obtain VIS particles. We also set up an experimental kit for collecting video. And create a dataset about particles moving under vibration. It includes 180 videos. We conducted training and testing experiments to prove the effectiveness of MaskTrack R-CNN on VIS particles. Below, we compare the effect of this final model with the baseline data. This is the first VIS application in civil engineering as far as we know in our own dataset. We believe the new application will innovate research ideas and provide a new direction for video awareness to the research community. We believe that this research will provide new ideas for microscopic analysis of particles.

## References

1. Linjie, Y., Yuchen, F., Ning, X.: Video instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5188–5197 (2019)
2. Kaiming, H., Georgia, G., Piotr, D., Ross G.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
3. Shaoqing, R., Kaiming, H., Ross, G., Jian. S.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99. (2015).
4. Jerónimo, P., Resende, R.: An assessment of contact and laser-based scanning of rock particles for railway ballast. *Transp. Geotech.* 100302 (2020)
5. Gao, G., Meguid, M.A., Chouinard, L.E., Xu, C.: Insights into the transport and fragmentation characteristics of earthquake-induced rock avalanche: numerical study. *Int. J. Geomech.* 04020157 (2020)
6. Liu, G.Y., Xu, W.J., Sun, Q.C., Govender, N.: Study on the particle breakage of ballast based on a GPU accelerated discrete element method. *Geosci. Front.* 461–471 (2020)
7. Bagherzadeh, H., Mansourpour, Z., Dabir, B.: Numerical analysis of asphaltene particles evolution and flocs morphology using DEM-CFD approach. *J. Petrol. Sci. Eng.* 108309 (2021)
8. Jiale, C., Rao Muhammad, A., Hisham, C., Fahad Shahbaz, K., Yanwei, P., Ling, S.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: ECCV (2020)
9. Jonathon, L., Idil Esen, Z., Bastian, L.: Unovost: Unsupervised offline video object segmentation and tracking. In: WACV (2020)
10. Ali, A., Sabarinath, M., Aljosa, O., Laura, L., Bastian, L.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: ECCV (2020)
11. Gedas, B., Lorenzo, T.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: CVPR (2020)
12. Yuan, L., Duan, Y.: A method of ore image segmentation based on deep learning. In: Proceedings of the International Conference on Intelligent Computing (ICIC), pp. 508–519 (2018)
13. Duan, J., Liu, X., Wu, X., Mao, C.: Detection and segmentation of iron ore green pellets in images using lightweight U-net deep learning network. *Neural Comput. Appl.* 1–16 (2019)
14. Liu, X., Zhang, Y., Jing, H., Wang, L., Zhao, S.: Ore image segmentation method using U-Net and Res\_Unet convolutional networks. *RSC Adv.* 9396–9406 (2020)
15. Lin, Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.: Efficient video object segmentation via network modulation. In: CVPR (2018)
18. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017)

# Infrared Dim Target Detection Based on Convolutional Neural Networks



Pinghuang Zhou and Wei Ai

**Abstract** Infrared dim target has small imaging area, less available features and target detection is highly susceptible to background clutter. Therefore, how to detect dim target accurately in complex scenes becomes a technical difficulty. In this paper, based on the powerful feature extraction ability of convolutional neural networks, a single frame infrared small target extraction networks is designed based on fully convolutional neural networks to extract the small target position accurately. Based on the sequence correlation, the energy accumulation in the space-time domain enhances the intensity of the dim target. Filter out the isolated noise through the adaptive filter and finally infrared dim target is extracted accurately. Experimental results show that Compared with traditional algorithms, this algorithm has the highest signal-to-noise ratio and signal-to-noise ratio gain, which is higher than 1 than the current practical Tophat algorithm, which proves that the proposed algorithm has higher extraction accuracy and lower false alarm rate.

**Keywords** Convolutional neural networks · Infrared dim target · Background suppression · Feature extraction

## 1 Introduction

Infrared small target detection has attracted more and more attention in recent years. However, due to the long detection distance, the small infrared target image imaging area and weak intensity, coupled with the interference of atmospheric clouds, make the infrared small target easy to be submerged in the background. In addition, the imaging process will produce noise, which is easy to be mistakenly detected as a target during detection, and the phenomenon of “high false alarm” appears [1, 2]. Therefore, how to accurately extract small and weak targets from infrared thermal images with low signal-to-noise ratio has become a major difficulty.

---

P. Zhou · W. Ai ()

Optics-Wuhan National Laboratory for Optoelectronics, Huazhong Institute of Electro-Optics, Wuhan, China

e-mail: [aiei@alumni.hust.edu.cn](mailto:aiei@alumni.hust.edu.cn)

In recent years, the research of infrared small target detection technology has developed rapidly, and many methods have been proposed. The methods are mainly divided into the following two categories: One is based on morphology [3], and spatial frequency domain filtering [4]; the other is Method based on feature detection [5, 6]. With the development of deep learning, various types of neural networks have emerged in endlessly, and have been applied to many computer vision tasks, such as face recognition [7], image classification [8], text detection and recognition and other fields. Compared with traditional algorithms, deep learning methods can greatly improve the accuracy of many tasks and improve the robustness of the algorithm. Many researchers have applied the idea of deep learning to the detection of small infrared targets [9–11].

This paper analyzes the characteristics of infrared small targets, studies the components of noise and background, and applies them to the design of convolutional neural networks based on the difference between target characteristics and noise characteristics, and proposes a convolutional neural network based on Gaussian properties of infrared small targets. The design idea is to use the powerful feature extraction capabilities of convolutional neural networks to solve the problem of small infrared targets with few traditional features and difficult to describe features. Through the feature extraction capabilities of convolutional neural networks, small infrared targets can be better extracted to achieve Infrared small target detection.

## 2 Infrared Image Feature Analysis

Generally speaking, infrared images are mainly composed of three parts: target, background and noise. The infrared small target image can generally be described by the mathematical model shown in the following formula:

$$f(i, j) = B(i, j) + T(i, j) + N(i, j) \quad (1)$$

where  $f(i, j)$  is the pixel value of the i-th row and the j-th column of the infrared image, which  $B(i, j)$ ,  $T(i, j)$ ,  $N(i, j)$  are the pixel values of the background, target, and noise at that pixel. Due to the optical diffraction effect, the target imaging on the infrared image is generally larger than the imaging size calculated by the geometric method. The energy of an ideal point target is dispersed during the imaging process, and the image appears as a light spot after diffraction, which is approximately Gaussian distribution. The gray distribution of the target area is mathematically described as follows:

$$f(x, y) = Ae^{-\frac{(x-x_0)^2+(y-y_0)^2}{\sigma^2}} \quad (2)$$

Among them,  $f$  represents the gray value of the small target,  $\sigma$  represents the actual size of the small target, and  $(x_0, y_0)$  is the position of the center point of the small target.

**Background characteristics:** The background is a relatively stable information component and basically does not change with the movement of the target. The background of the infrared image is mainly composed of the low-frequency part of the cloud layer with high correlation. The background of the infrared image studied by the airborne weak and small target detection system is mainly the sky and the cloud layer. Due to physical laws, the cloud layer distribution is continuous and has Greater connectivity.

**Noise characteristics:** Infrared images contain different types of noise. From the analysis of the components of the infrared imaging system, the entire system includes the optical system, the circuit system, the scanning system and the detector, and each component will generate noise. The noise roughly conforms to the Gaussian distribution, which is a Gaussian mixture model distribution.

### 3 Single-Frame Infrared Small Target Extraction Network Design

Through the analysis in Chap. 2, the intensity distribution of a single frame infrared small target conforms to the Gaussian function, and the characteristics of the background and noise are mainly low-frequency and Gaussian mixed distribution. Accordingly, a single-frame infrared small target extraction network can be designed. It is based on a fully convolutional network, which can handle pictures of different sizes. In addition, considering more local characteristics, the network does not include a fully connected layer.

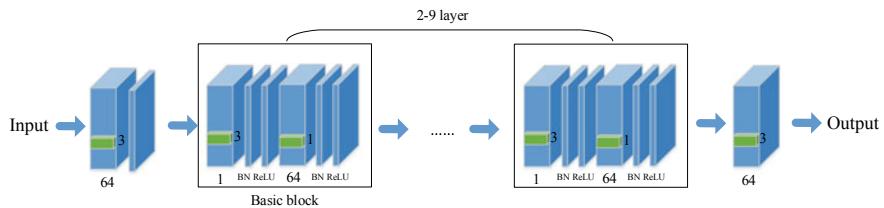
#### 3.1 The Main Body Design of the Network

The main structure of the network uses the DnCNN network [12] as the base network, and mainly uses the convolutional layer to extract the characteristics of noise and the ReLU activation function to perform the operation of non-linear transformation. Its main structure is shown in Table 1.

Reduction in the amount of parameters: The second to 9th layers adopt the MobileNet method, and use deep separable convolution for training. When performing depthwise convolution, only one-dimensional  $3 \times 3$  convolution is used to extract the features of the convolution kernel of the input channel (without feature combination), and only the dimension  $1 \times 1$  of the output result is used when performing pointwise convolution. The convolution kernel performs feature combination. This can reduce the amount of parameters. If  $3 \times 3$  convolution kernel + BN

**Table 1** Network structure composition

Number of layers	Specific composition convolution	kernel size	Number of channels
The first layer	Conv1_1 + Relu	$3 \times 3$	64
The second layer	Con2_1/DW + BN + Relu	$3 \times 3$	1
	Con2_2/PW + BN + Relu	$1 \times 1$	64
...	...	...	...
Tenth layer	Conv10_1	$3 \times 3$	64
Output			1

**Fig. 1** Network structure diagram

+ ReLU is used directly, the parameter amount for each layer is  $3 \times 3 \times 64 \times 64$ . If the depth separation convolution method is adopted, the parameter amount of each layer is  $3 \times 3 \times 64 + 1 \times 1 \times 64 \times 64$ , which is about 1/3 of the original, which greatly reduces the parameter amount. The final network structure diagram is shown in the figure: (see Fig. 1).

### 3.2 Network Training

The training of the network uses the error back propagation algorithm to train the network, and the error learning uses the residual learning method [13, 14]. Among them, the gradient descent algorithm selects Adam algorithm [15] for bias correction, and the correction uses the origin-initialized first-order moment (momentum term) and (non-central) second-order moment estimation, which can prevent the optimization process from entering the local optimal solution.. The batch size in pre-training is selected as 128, the number of iterations is 50, and the learning rate is 0.001. After 30 times, the learning rate becomes 0.1 before every 5 rounds. In post-training, the batch size is selected as the size of all artificially labeled data sets, the number of iterations is 100, the learning rate is the final learning rate of the pre-training, and the learning rate becomes 0.1 before every 20 rounds.

## 4 Weak Target Detection in Sequence Images

Considering that in the process of extracting the star target from a single frame image, there will still be some isolated noises. These noises are similar to the star target characteristics, but for the sequence, the noise is generally randomly distributed or moving fast, and the star target is in the image. The movement is slow in the middle, so it is necessary to use the correlation in the time domain to eliminate the noise and further accurately obtain the star target.

For the N frames of images output through the network, first, a spatial energy superposition is performed for each frame of the image to increase the intensity of the point target. Since the size of the point target is small, the method of accumulating the window is adopted to enhance the target energy. The size of the window is  $3 \times 3$  or  $5 \times 5$ . For the pixels with gray values greater than 10 in the filtered image, the sum is performed as following.

$$f(x, y) = \sum_{i=-w}^w \sum_{j=-w}^w f(x + i, y + j) \quad (3)$$

After the image is superimposed in the space, the energy of the point target is strengthened, and the multi-frame correlation of the point target and the random distribution of noise are used to accumulate the energy of the point target in the time domain. For the selected N frames of images, the superposition operation is as the calculation formula:

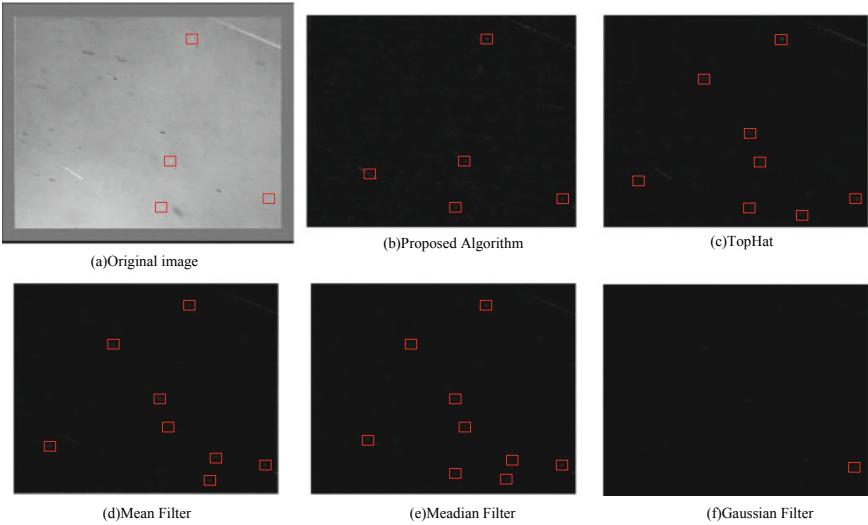
$$f(i, j) = B(i, j) + T(i, j) + N(i, j) \quad (4)$$

For the final overlay image, construct an adaptive filter, input the new output image into the filter, filter out noise, output the target position of the point, and mark it in each image.

## 5 Experimental Result

Figure 2a is the input image, which includes objects, sky, and a relatively uniform background. (b) is the result of the algorithm proposed in this paper, and (c-f) are the results of TopHat algorithm, mean filter, median filter, and Gaussian filter respectively. It can be seen that the algorithm proposed in this paper accurately detects the position of the faint star target.

In order to evaluate the performance of the algorithm more clearly, the image signal-to-noise ratio SNR and the signal-to-noise ratio gain coefficient GSNR are used to compare the performance of the algorithm in the above figure. The SNR calculation formula is:



**Fig. 2** Comparison of different algorithm results

$$SNR = \frac{m_T - m_B}{\sigma_B} \quad (5)$$

Among them,  $m_T$  is the average gray value of the target area in the image,  $m_B$  is the average gray value of the background area of the image, and  $\sigma_B$  is the standard deviation of the background area. The larger the signal-to-noise ratio SNR, the more prominent the target in the image, the higher the discrimination between the target and the background, and the less difficult the detection. The calculation formula of GSNR is as follows:

$$GSNR = \frac{SNR_0}{SNR_1} \quad (6)$$

Among them,  $SNR_0$  is the signal-to-noise ratio of the original image, and  $SNR_1$  is the signal-to-noise ratio of the image processed by the algorithm. The larger the signal-to-noise ratio gain coefficient, the better the performance of the algorithm for target enhancement. The results are shown in Table 2.

**Table 2** Comparison results of different algorithms

	Original image	TopHat	Mean filter	Median filter	Gaussian filter	Proposed algorithm
SNR	1.132	8.467	7.091	8.141	7.045	9.504
GSNR	1	7.480	6.264	7.192	6.223	8.396

It can be seen that whether from the visual detection results or the signal-to-noise ratio data comparison, the detection results of this paper are more accurate, which further proves the effectiveness of the algorithm in this paper.

## 6 Conclusion

In this paper, by analyzing the characteristics of small infrared targets and the intensity and distribution characteristics of background and noise, a single-frame infrared small target extraction network is designed based on a fully convolutional neural network, and the network is properly trained to achieve a relatively accurate extraction of small targets from single-frame images. Position, and then use the sequence correlation of small targets to accumulate energy in the time and space domain to enhance the strength of weak targets, and design an adaptive filter to filter out isolated noises, and finally accurately extract small and weak infrared targets. Experimental results prove that the results of this algorithm are better than other traditional algorithms.

## References

1. Huakai L.: Research on Aerial Infrared Small Target Detection Algorithm under Complex Cloud Background. Beijing University of Technology (2019)
2. Dehghani A., Pourmohammad A.: Small target detection and tracking based on the background elimination and Kalman filter. In: International Symposium on Artificial Intelligence and Signal Processing, pp. 328–333 (2015)
3. Peng, D.: Infrared small target detection method based on morphology and facet kernel filtering. J. Kashgar Univ. **40**(06), 60–63 (2019)
4. Jie, H., Yunhong, X.: Infrared small target detection method based on high-pass filtering and image enhancement. Infrared Technol. **35**(5), 279–284 (2013)
5. Tao, W., Wenzhong, H., Xiaolu, C.: Single-frame infrared small target detection algorithm based on local features. Laser Infrared **46**(3), 368–371 (2016)
6. Min, W., Wenqing, X.: Infrared small target tracking algorithm based on multi-features and evaluation models. J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.) **45**(10), 1–6 (2017)
7. Luning, W.: Research on Face Detection Algorithm Based on Fully Convolutional Neural Network. Zhejiang University (2017)
8. Mingwei, L.: Research on Convolutional Neural Network Method in Image Classification. Nanjing University of Posts and Telecommunications (2016)
9. Lin, L., Wang, S., Tang, Z.: Using deep learning to detect small targets in infrared oversampling images. J. Syst. Eng. Electron. **29**(5), 947–952 (2018)
10. Wang, W., Qin, H., Cheng, W. et al.: Small target detection in infrared image using convolutional neural networks. Proc. Opt. Sens. Imaging Technol. Appl. 2017.1046250(24)
11. Shuangchen, W., Zhengrong, Z.: Infrared small target detection based on deep convolutional neural network. J. Infrared Millimeter Waves **38**(03), 371–380 (2019)
12. Zhang, K., Wangmeng, Z., Yunjin, C., et al.: Beyond a Gaussian Denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
13. Wu, C., Chen, X., Ji, D., et al.: Image denoising combining deep residual learning and perceptual loss. J. Image Graph. **23**(10), 1483–1491 (2018)

14. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. **2016**, 770–778 (2016)
15. Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. Adv. Neural Inf. Process. Syst., pp. 2377–2385 (2015)

# Improvement of Multi Frequency Heterodyne Phase Unwrapping in Extreme Environment



Bingwei Zhang , Junyi Lin , Shaoning Lin , Yabin Liu , and Kaiyong Jiang

**Abstract** Multi-frequency heterodyne (MFH) phase unwrapping algorithm plays an important role in 3D surface measurement because of its high precision and anti-interference features. However, under extreme environments such as over-dark or over-bright or low signal-to-noise ratio (SNR), the phase unwrapping may fail due to the increase of phase error. In order to improve the robustness of algorithm, we proposed an improved MFH algorithm to reduce the error accumulation of synthetic phase in the process of MFH algorithm. We expanding synthetic phase using the rounding function to remove the error of the synthesis phase, and then scaling down expanded synthetic phase to obtain the synthetic phase in  $[0, 2\pi]$ , which has higher accuracy than before. The simulation and experimental results show that, in the same conditions, our method is more robust than traditional MFH algorithms. The proposed method is applied to the surface measurement of relay components, and it can effectively avoid the failure of phase unwrapping in the actual measurement.

**Keywords** Multi-frequency heterodyne · Phase error · Tolerance

## 1 Introduction

Fringe projection profilometry (FPP) is an important three-dimensional measurement method, which has the advantages of non-contact, low cost and high precision, and is widely used in various fields [1–3]. The method uses the obtained phase information for phase matching to obtain the distance between two matching points and convert it into point cloud data to complete the measurement of 3D surface. In order to obtain the pre-coded phase value, the phase-shift method is one of the most commonly used and high precision phase measurement methods at present, but it can only obtain

---

B. Zhang · J. Lin · S. Lin · Y. Liu · K. Jiang ()

Fujian Key Laboratory of Special Energy Manufacturing, Huaqiao University, Xiamen 361021, China

e-mail: [jiangky@hqu.edu.cn](mailto:jiangky@hqu.edu.cn)

Xiamen Key Laboratory of Digital Vision Measurement, Huaqiao University, Xiamen 361021, China

the wrapping phase in the range of  $[-\pi, \pi]$ . In order to meet the requirement of continuous phase information of the full field, phase unwrapping algorithm should be used to recover the absolute phase maps from the wrapped phase [4, 5].

At present, among the methods of phase unwrapping, the MFH algorithm has been used by many researchers and companies [6–12] because of its high measurement accuracy and fast speed. However, in practical application, when the measurement environment is over-dark or over-bright or low SNR [6] the phase unwrapping may fail due to the increase of phase error. Chen [7] and Huang [8] analyzed the field of phase error points after unwrapping, and corrected the errors, but it is easy to overcorrection for objects with discontinuous surfaces. Zhang [9] avoiding the effects of phase errors by filtering the lower frequency fringe pattern. Lei [10] corrected the phase after unwrapping by improving the multi-frequency heterodyne principle, but this method is only effective for dual-frequency heterodyne. Zhao [11] improved the existing MFH method, avoiding the occurrence of phase jump, but it needs some constraints on the frequency combination. Lai [12] doubles the tolerance of phase error by projecting the synthetic patterns to the object directly. However, in the above phase unwrapping algorithms, the jump error cannot be completely eliminated, and the unwrapping failure may still occur when the phase error is too large.

Therefore, in order to further improve the robustness and the tolerance of phase error of MFH in the phase unwrapping process, this paper analyzes the error of the traditional MFH phase unwrapping method. According to the analyzed error propagation relationship, the formula is improved, and the error accumulation of the synthetic phase is narrowed by using rounding and scaling down transformation. Thus, the fault tolerance ability of phase unwrapping method is improved.

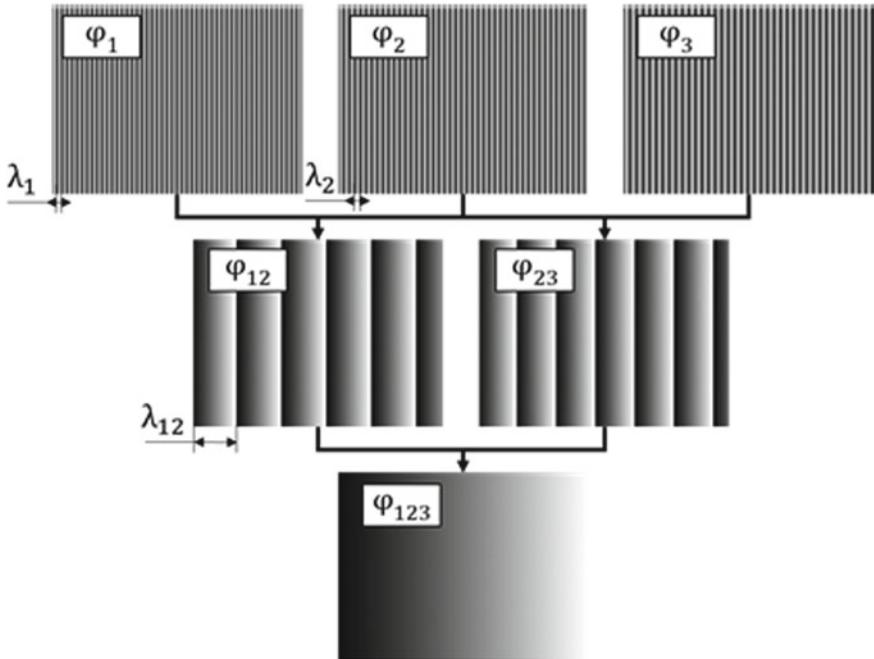
## 2 Principle

### 2.1 Multi-frequency Heterodyne

In phase matching, in order to meet the requirements of global continuous phase information, the phase shift method is adopted to solve the unwrapping phase in the range of  $[0, 2\pi]$ , and then it needs to recover the absolute phase [13]. The multi-frequency heterodyne phase unwrapping algorithm is widely used due to its high precision and high speed. Its principle is shown in Fig. 1.

As shown in Fig. 1, the  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_{12}$  are periods of the phase  $\varphi_1$ ,  $\varphi_2$  and  $\varphi_{12}$  respectively, and  $\lambda_1 < \lambda_2$ . By projecting phase pattern  $\varphi_1$ ,  $\varphi_2$  and  $\varphi_3$  with periods  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  respectively, we can obtain a larger synthetic phase  $\varphi_{12}$  and  $\varphi_{23}$  by Eq. (1), and then obtain  $\varphi_{123}$  by  $\varphi_{12}$  and  $\varphi_{23}$  in the same way.

$$\varphi_{mn} = \begin{cases} \varphi_m - \varphi_n, & \varphi_m \geq \varphi_n \\ 2\pi + \varphi_m - \varphi_n, & \varphi_m \leq \varphi_n \end{cases}, \quad (m, n = 1, 2; 2, 3) \quad (1)$$



**Fig. 1** Illustration of multi-frequency heterodyne principle

In order to satisfy the need for obtaining full-field continuous phase, an appropriate final period  $\lambda_{123}$  must be chosen so that the synthetic phase  $\varphi_{123}$  can cover the entire measurement range, where the following relationship between initial phase period and synthetic phase period is satisfied:

$$\lambda_{mn} = \left| \frac{\lambda_m \lambda_n}{\lambda_m - \lambda_n} \right|, (m, n = 1, 2; 2, 3) \quad (2)$$

In order to unwrap the phase  $\varphi_1$  or  $\varphi_2$  into absolute phase. The unwrapped absolute phase of the fringed pattern can be described as follows:

$$\Phi_m(x) = \varphi_m(x) + k_m(x) \times 2\pi, \quad (m = 1, 2, 3) \quad (3)$$

In this case,  $k_m(x)$  represents the order of the phase. After calculating the synthetic phase  $\varphi_{123}$  by  $\varphi_{12}$  and  $\varphi_{23}$ , which uses the same way as Eq. (1). The order  $k_m(x)$  can be obtained by the following formula [13]. The round function is used to find the nearest integer, so as to estimate the fringe order.

$$k_m(x) = \text{ROUND} \left( \frac{\varphi_{123}(x)(\lambda_{123}/\lambda_m) - \varphi_m(x)}{2\pi} \right) \quad (4)$$

According to the above discussion, the heterodyne principle can be used to obtain the synthetic phase  $\varphi_{123}$  covering the entire measurement range, which is used to unwrapping the absolute phase. However, in our actual measurement process, it is found that the error accumulation during the synthetic phase process may lead to the failure of phase unwrapping.

## 2.2 Phase Unwrapping Error Analysis

In order to further explore the reason of the failure of the traditional MFH phase unwrapping algorithm, it is necessary to analyze the phase error propagation in MFH phase unwrapping algorithm. Firstly, we assume that the phase error calculated by the phase-shift method is as follows:

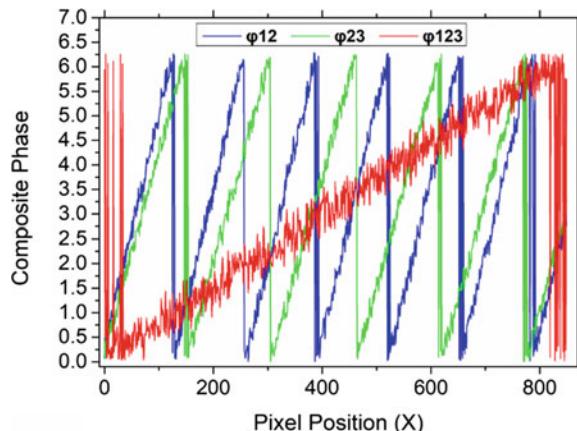
$$\varphi_{real} = \varphi_{theory} + \Delta\varphi \quad (5)$$

$\varphi_{real}$  is the actual wrapped phase,  $\varphi_{theory}$  is the theoretical wrapped phase, and  $\Delta\varphi$  is the phase error generated by the nonlinearity of the projector and CCD camera, random noise, phase shift and other errors from the system itself [14, 15], we suppose the phase errors are independent of the frequency of the fringe pattern. So in the fringe pattern of different frequencies we can get the following relation [12]:

$$\Delta\varphi_1(x) = \Delta\varphi_2(x) = \Delta\varphi_3(x) \quad (6)$$

In the process of phase unwrapping, according to Eq. (1), since the synthetic phase is formed by two wrapped phases, the phase error will also accumulate with each other, as shown in Fig. 2.

**Fig. 2** Illustration of synthesized phase error accumulation



Since  $\Delta\varphi_{12}$  is formed by the  $\Delta\varphi_1$  and  $\Delta\varphi_2$ , according to Eq. (6), the value of  $\Delta\varphi_{12}$  and  $\Delta\varphi_{23}$  is double  $\Delta\varphi$ .

$$\Delta\varphi_{12}(x) = \Delta\varphi_{23}(x) = 2\Delta\varphi(x) \quad (7)$$

And then,  $\varphi_{12}$  and  $\varphi_{23}$  will be further to synthesize  $\varphi_{123}$ , which accumulation the phase error again as quadruple  $\Delta\varphi$ .

$$\Delta\varphi_{123}(x) = \Delta\varphi_{12}(x) + \Delta\varphi_{23}(x) = 4\Delta\varphi(x) \quad (8)$$

It can be known from Eq. (8) that the synthetic phase  $\varphi_{123}$  is synthesized from  $\varphi_1$ ,  $\varphi_2$  and  $\varphi_3$ , which has a large phase error. According to Eq. (4), the synthetic phase is multiplied by the coefficient  $\lambda_{123}/\lambda_m$ , and the error is further enlarged. Under the integer function *Round()*, phase jump will occur when the sum of errors exceed 0.5, which can be described as follows:

$$\left| \frac{4\Delta\varphi(x)V_o + \Delta\varphi(x)}{2\pi} \right| < 0.5 \quad (9)$$

$$V_o = \lambda_{123}/\lambda_m \quad (m = 1, 2, 3) \quad (10)$$

In this case,  $V_0$  represents the ratio of the final synthetic phase period to the initial phase period. According to Eq. (9), it can be seen that the unwrapping quality of traditional MFH algorithm is related to the value of  $V_0$  and  $\Delta\varphi$ . However, due to the influence of illumination, jitter and other factors in the actual measurement, the generation of phase error  $\Delta\varphi$  is difficult to avoid. In addition, the value of  $V_0$  is relatively large. For example  $\lambda_1 = 16, \lambda_2 = 18, \lambda_3 = 21$ , the value of  $V_0$  is going to be 63. It will significantly enlarge the influence of  $\Delta\varphi$ , which makes it difficult to keep the absolute value of errors in Eq. (9) below 0.5. Therefore, we can imagine that the traditional MFH method has the problem of weak anti-interference ability in the actual measurement.

### 2.3 Improvement of MFH Algorithm

In order to reduce the influence of phase error on phase unwrapping, we propose an improved method, which can reduce the secondary accumulation of the phase error. After the synthetic phase is calculated using the heterodyne principle in Eq. (1), we can narrow the phase error of synthetic phase by using the method we proposed. In proposed method, the synthetic phase is first expanded by Eq. (11), which use the round function to calculate the fringe order and remove the error of the synthesis phase. And then the expanded phase is scaled down by Eq. (12), to obtain the synthetic phase in  $[0, 2\pi]$ . After the above two steps, the higher precision of the synthetic phase will be obtained, as shown in Eq. (13). It provides a reliable guarantee for the next

step phase unwrapping, by narrow the phase error of synthetic phase. The calculation formula of the above two steps are as follows:

$$\Phi_m(x) = \varphi_m(x) + 2\pi \times \text{ROUND}\left(\frac{\varphi_{mn}(x)(\lambda_{mn}/\lambda_m) - \varphi_m(x)}{2\pi}\right) \quad (11)$$

$$\varphi_{mn}(x) = \Phi_m(x)(\lambda_m/\lambda_{mn}), \lambda_m/\lambda_{mn} < 1 \quad (12)$$

According to Eqs. (11) and (12), we can obtain the synthetic phase  $\varphi_{mn}^{re}(x)$  with reduced phase error. The formula is as follows:

$$\varphi_{mn}^{re}(x) = \frac{\lambda_m \left[ \varphi_m(x) + 2\pi \times \text{ROUND}\left(\frac{\varphi_{mn}(x)(\lambda_{mn}/\lambda_m) - \varphi_m(x)}{2\pi}\right) \right]}{\lambda_{mn}} \quad (13)$$

In the above equation,  $\lambda_m$  represents the main phase period;  $\lambda_{mn}$  represents the synthetic phase period, and  $\varphi_{mn}(x)$  represents the synthesis phase.  $\varphi_{mn}^{re}(x)$  represents the synthetic phase with narrowed error, which is expanded by main phase and synthetic phase and then be scaled down to  $[0, 2\pi]$ , as shown in Eq. (13). Through error analysis of the above method, it can be seen that the accumulation error of  $\varphi_{mn}(x)$  has been cut off after rounding function of phase unwrapping. In the new synthesis phase  $\varphi_{mn}^{re}(x)$ , only the error is generated by the main phase  $\varphi_m(x)$  remain. After that the synthesis phase is divided by the coefficient  $\lambda_{mn}/\lambda_m > 1$ , so the error of the synthetic phase will be further narrowed. From this, we can draw the following conclusion:

$$\Delta\varphi_{mn}^{re}(x) < \Delta\varphi_m(x) < \Delta\varphi_{mn}(x) \quad (14)$$

$$\varphi_{123}^{re} = \begin{cases} \varphi_{12}^{re} - \varphi_{23}^{re} & \varphi_{12}^{re} \geq \varphi_{23}^{re} \\ 2\pi + \varphi_{12}^{re} - \varphi_{23}^{re} & \varphi_{12}^{re} \leq \varphi_{23}^{re} \end{cases} \quad (15)$$

According to Eq. (15), after rounding and scaling down, the error of the synthetic phase  $\varphi_{123}^{re}(x)$  is narrowed compared to the original synthetic phase  $\varphi_{123}$ . However, in the process of phase unwrapping, if the main phase  $\varphi_m$  is directly unwrapping by the synthetic phase  $\varphi_{123}^{re}(x)$ , the phase unwrapping may fail because the  $V_0$  coefficient is large. In order to avoid the influence of  $V_0$ , we rewrite Eq. (4) as Eqs. (16) and (17). First we use  $\varphi_{123}^{re}(x)$  to unwrapping the synthetic phase  $\varphi_{mn}^{re}(x)$ , as shown in Eq. (16), replaces  $V_0$  with a smaller coefficients  $\lambda_{123}/\lambda_{mn}$ . Then the use Eq. (17) to unwrapping the main phase  $\varphi_m$ . With these improvements, both the phase error and the proportional coefficient are reduced.

$$k_{mn}^{re}(x) = \text{ROUND}\left(\frac{\varphi_{123}^{re}(x)(\lambda_{123}/\lambda_{mn}) - \varphi_{mn}^{re}(x)}{2\pi}\right) \quad (16)$$

$$k_m(x) = \text{ROUND} \left( \frac{(\varphi_{mn}^{re}(x) + 2\pi k_{mn}^{re}(x))(\lambda_{mn}/\lambda_m) - \varphi_m(x)}{2\pi} \right) \quad (17)$$

Following, we will analyze the error of the newly proposed method. First of all, in the process of getting the synthetic phase  $\varphi_{mn}^{re}(x)$ , according to Eq. (13), due to the existence of the rounding function, the phase error shall meet the following requirements:

$$\left| \frac{2\Delta\varphi(x) \times \lambda_{mn}/\lambda_m + \Delta\varphi(x)}{2\pi} \right| < 0.5 \quad (18)$$

As shown in Eq. (18), since the main phase is not affected by error superposition, and the value of  $\lambda_{mn}/\lambda_m$  is relatively small compared with  $\lambda_{123}/\lambda_m$ , so phase unwrapping failure is not easy to happen in Eq. (18).

$$\left| \frac{\Delta\varphi(x) \times V_1 + \Delta\varphi(x) \times V_2 + \Delta\varphi(x) \times V_3}{2\pi} \right| < 0.5 \quad (19)$$

$$\left| \frac{\Delta\varphi(x) \times V_4 + \Delta\varphi(x)}{2\pi} \right| < 0.5 \quad (20)$$

where

$$\begin{cases} V_1 = \frac{\lambda_{123}/\lambda_{mn}}{\lambda_{12}/\lambda_1} < V_0 \\ V_2 = \frac{\lambda_{123}/\lambda_{mn}}{\lambda_{12}/\lambda_2} < V_0 \\ V_3 = \frac{1}{\lambda_{12}/\lambda_1} < 1 \\ V_4 = \frac{\lambda_{mn}/\lambda_m}{\lambda_{12}/\lambda_1} \end{cases} \quad (21)$$

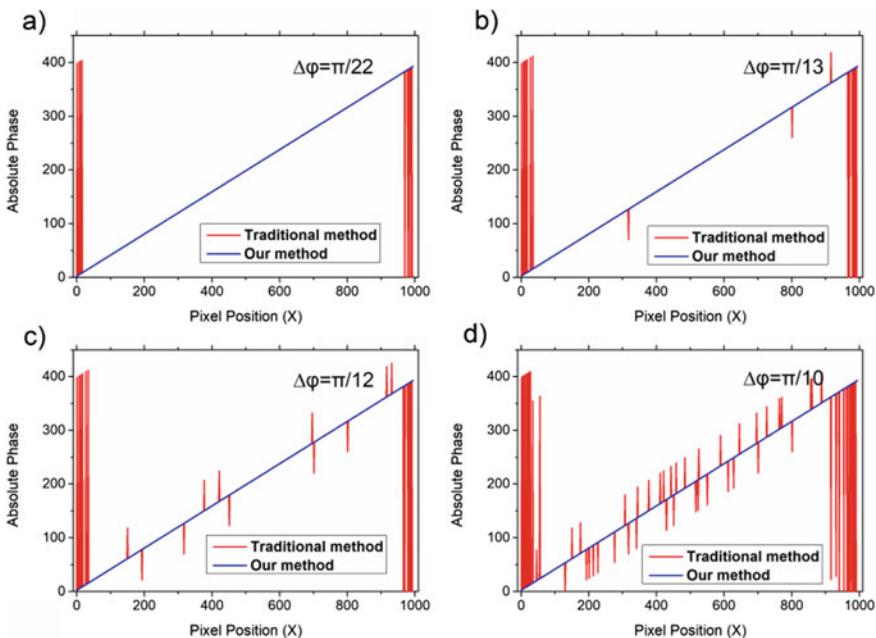
In the final step of phase unwrapping using the synthetic phase  $\varphi_{123}^{re}(x)$ , we converted the phase unwrapping process into two steps as shown in Eqs. (16) and (17) to reduce the phase error, according to the above formula, we can get that the sum of phase error needs to meet the requirements of Eqs. (19) and (20) at the same time.

According to Eq. (21), it can be seen that all the coefficients in the proposed method are smaller than the traditional method. Therefore, it can be concluded theoretically that the proposed method can suppress the influence of phase error better than the traditional method.

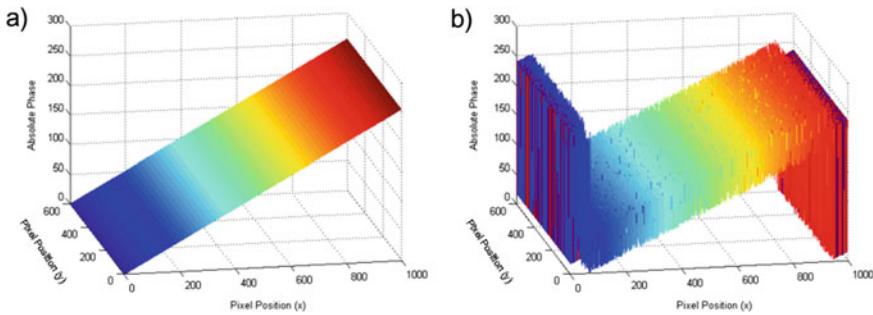
### 3 Experiments and Simulations

#### 3.1 Simulation

To verify the feasibility of the algorithm presented in this paper, fringe pitch with ( $\lambda_1 = 16$ ,  $\lambda_2 = 18$ ,  $\lambda_3 = 21$ ) is selected for experimental simulation. And we added random noise with a range of  $[-\Delta\varphi, \Delta\varphi]$  into the fringe pattern to simulate the generation of nonlinear phase errors under real environments. Then the proposed method and the traditional method are used for phase unwrapping respectively. When  $\Delta\varphi = \pi/22$ , both algorithms can correctly unwrapping the phase, in Fig. 3a. And the absolute phase of each algorithms matches perfectly, which means the proposed method is correct and effective. However, when  $\Delta\varphi > \pi/13$ , there are  $2\pi$  jumps for the traditional method, in Fig. 3b–d. The proposed method is unwrapping successfully even  $\Delta\varphi = \pi/10$ , in Figs. 3d and 4. The simulation results agree well with the theoretical analysis, which means the proposed method should be more robust in practice.



**Fig. 3** Results of phase unwrapping with different conditions. **a**  $\Delta\varphi = \pi/22$ . **b**  $\Delta\varphi = \pi/13$ . **c**  $\Delta\varphi = \pi/12$ . **d**  $\Delta\varphi = \pi/10$



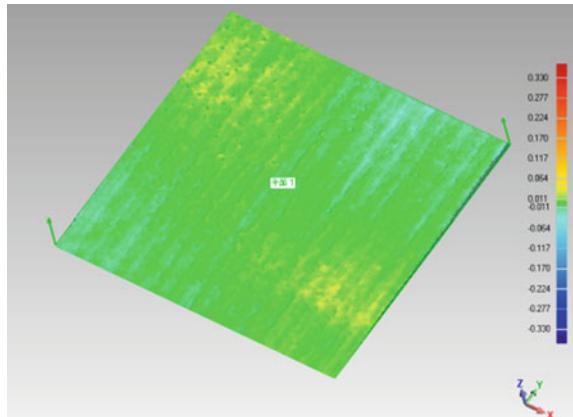
**Fig. 4** Absolute phase unwrapping result with phase error of  $\pi/10$ . **a** Phase unwrapping in proposed method. **b** Phase unwrapping in traditional methods

### 3.2 Experiments

In order to further evaluate the feasibility of the algorithm, we developed a fringe projection measurement system consisting of a PC, a projector (DLP4500, TI) and CCD cameras (The Imaging Source DMK33UX174 with resolution 1920 \* 1200) to verify our method. First we project a stripe pattern of period ( $\lambda_1 = 16$ ,  $\lambda_2 = 18$ ,  $\lambda_3 = 21$ ) onto a white standard plate. The algorithm proposed in this paper is used to phase unwrapping, and the reconstruction results are shown in Fig. 5. The reconstructed plane surface is smooth and the point cloud is evenly distributed. We carried out plane fitting for point cloud data, and analyze the error as shown in Table 1.

In addition, in order to test the robustness of the algorithm in the actual measurement, we apply it to the surface measurement of relay components. Since the surface material of relay components is composed of metal and plastic, the low reflectivity of the metal surface will lead to the decline of the SNR. Moreover, the random error

**Fig. 5** Reconstruction results of plane surface

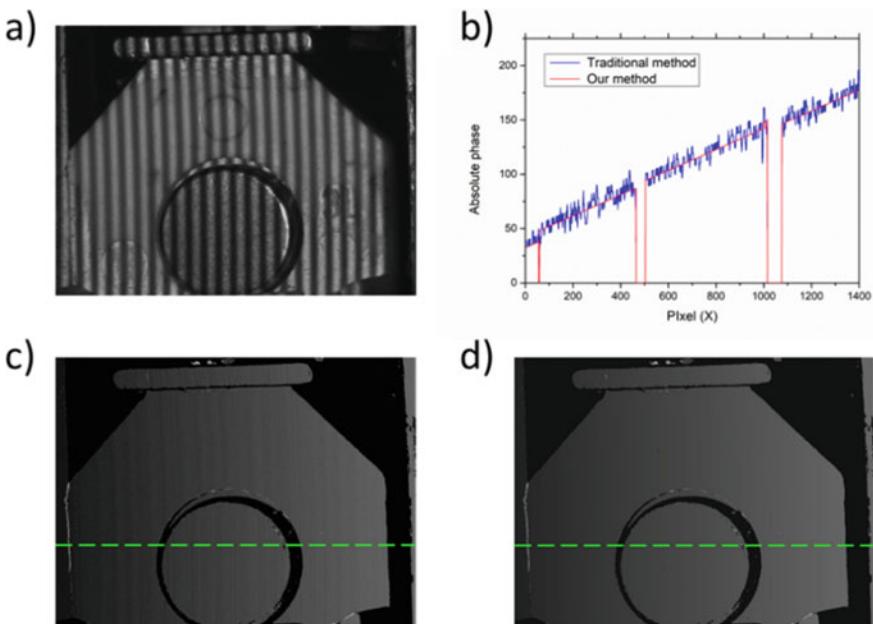


**Table 1** Error analysis of plane point cloud data

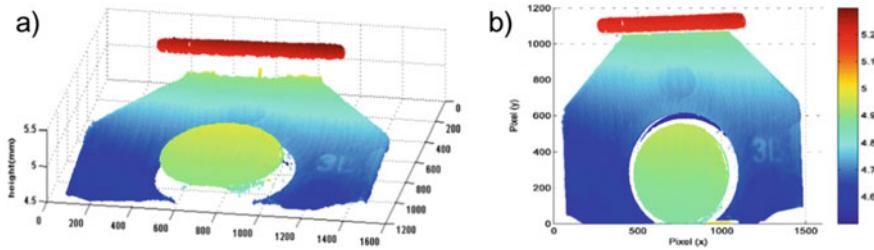
Index of plane point cloud data	Maximum deviation (mm)	Mean deviation (mm)	Standard deviation (mm)
1	0.0368	0.009	0.0111
2	0.0345	0.015	0.0154
3	0.0342	0.012	0.0143

caused by the reflection of the microstructure on the metal surface will further aggravate the degree of phase error. If the traditional MFH phase unwrapping method is directly used, the phase unwrapping will fail, as shown in Fig. 6c. The algorithm proposed in this paper can successfully unwrap the phase even if there are phase errors mentioned above. As can be seen from Fig. 6b-d, compared with the traditional method, the proposed method has better robustness and can well suppress the jump error caused by phase error.

After 3D reconstruction of the full-field phase, the results as shown in Fig. 7 can be obtained. It can be seen that the fine lines and protrusions on the surface of relay components have been well reconstructed, and the surface is smooth and delicate.



**Fig. 6** The relay components and the results of phase unwrapping. **a** The captured image. **b** The absolute phases of green line in the next two pictures. **c** Phase unwrapping in traditional method. **d** Phase unwrapping in proposed method



**Fig. 7** Reconstruction results of relay components. **a** Side view. **b** Top view

## 4 Conclusions

In the traditional MFH algorithm phase unwrapping, when the measurement environment is over-dark or over-bright or the SNR is low, phase unwrapping may fail due to the phase error. The proposed improved MFH algorithm reduce the error accumulation of synthetic phase to increase the robustness of the unwrapping algorithm. It uses the round function to calculate the fringe order and remove the error of synthesis phase, and scaling down to obtain the synthetic phase in  $[0, 2\pi]$ , which provides a reliable guarantee for the next step phase unwrapping.

The experiment and simulation analysis verify that the algorithm proposed in this paper can unwrap the phase in extreme environments. Compared with the traditional algorithm, the proposed algorithm has stronger tolerance of the phase error. However, the jump error still occurs at both ends of the full-field phase, in order to further ensure the success of phase unwrapping, we should also try to avoid the positions of both ends in the actual measurement.

**Acknowledgements** This work was financially supported by Fujian Province Industry-University-Research Program (2019H6016).

## References

1. Chen, F., Brown, G.M., Song, M.: Overview of three-dimensional shape measurement using optical methods. *Opt. Eng.* **39**(1) (1999)
2. Salvi, J., Fernandez, S., Pribanic, T., et al.: A state of the art in structured light patterns for surface profilometry. *Pattern Recogn.* **43**(8), 2666–2680 (2010)
3. Jalkio, J.A., Kim, R.C., Case, S.K.: Three dimensional inspection using multistripe structured light. *Opt. Eng.* **24**(6), 966–974 (1985)
4. Reich, C.: 3-D shape measurement of complex objects by combining photogrammetry and fringe projection. *Opt. Eng.* **39**(1), 224–231 (2000)
5. Zhang S.: Absolute phase retrieval methods for digital fringe projection profilometry: a review. *Opt. Lasers Eng.*, **107**(August), 28–37 (2018)
6. Chen, S., Xia, R., Zhao J., et al.: Analysis and reduction of phase errors caused by nonuniform surface reflectivity in a phase-shifting measurement system. *Opt. Eng.* **56**(3), 033102 (2017)

7. Chen, L., Deng, W., Lou, X.: Phase unwrapping method base on multi-frequency interferometry. *Opt. Tech.* **38**(1), 73–78 (2012)
8. Huang, Y., Lou, X.: Phase correction and matching based on multi-frequency heterodyne method. *J. Appl. Opt.* **35**(2), 237–241 (2014)
9. Zhang S.: Phase unwrapping error reduction framework for a multiple-wavelength phase-shifting algorithm. *Opt. Eng.* **48**(10), 105601 (2009)
10. Lei, Z., Li, J.: Full automatic phase unwrapping method based on projected double spatial frequency fringes. *Acta Optica Sinica* **26**(1), 39–42 (2006)
11. Chen, S., Zhao, J., Xia, R.: Improvement of the phase unwrapping method based on multi-frequency heterodyne principle. *Acta Optica Sinica* **36**, 409(04): 155–165 (2016)
12. Lai, J., Li, J., He, C., et al.: A robust and effective phase-shift fringe projection profilometry method for the extreme intensity. *OPTIK-STUTTGART* (2019)
13. Zuo, C., Huang, L., Zhang, M., et al.: Temporal phase unwrapping algorithms for fringe projection profilometry: a comparative review. *Opt. Lasers Eng.* **85**(October), 84–103 (2016)
14. Zhang, C., Zhao, H., Jiang, K.: Fringe-period selection for a multi frequency fringe-projection phase unwrapping method. *Measurementence Technol.* **27**(8): 085204 (2016)
15. Pan, B., Qian, K., Huang, L., et al.: Phase error analysis and compensation for nonsinusoidal waveforms in phase-shifting digital fringe projection profilometry. *Opt. Lett.* **34**(4), 8–416 (2009)

# A Quick and Accurate Method to Identify Betel Nut Based on Mobilenetv3



Yun Dai Ming Lu and Zuguo Chen

**Abstract** The rapid development of artificial intelligence has brought a lot of convenience to our life and also affected the development of industry. Betel nut is a very popular food in China, which is fall with the majority of men. The market demand is also expanding. Combining the process of betel nut with machine vision is a great attempt to improve the efficiency of the plant. At present, the processing of betel nut is complex and there is an urgent need to improve the degree of automation. The development of machine vision can provide a good idea. In this paper, a lightweight network model is used to solve the classification problem of betel nut core in view of the key process of betel nut processing. At the same time, through comparative experiments, this paper selects the best performance indicators and the best recognition of mobilenetv3 network as an accurate identification method.

**Keywords** Artificial intelligence · Machine vision · Betel nut · Mobilenetv3

## 1 Introduction

Betel nut is a fast-consuming food popular in southern China, and its processing process is relatively complex. The key process of betel nut processing: cutting seeds, denuclearization and halogenation. Due to the irregular shape of betel nut, a large number of manual classifications are still required and the workload is huge. In order to reduce labor costs and improve production efficiency. This paper will adopt the lightweight network model, from the point of view of machine vision, to improve the degree of automation in the industry to provide new ideas.

At present, the domestic research on betel nut is still only in infancy time. Xu et al. [1] extracted the color features, shape and texture characteristics of betel nut to grade the varieties, and replaced them with vector machines, and its correct recognition rate is more than 90.38%. Zhu et al. [2] propose semantic segmentation method to detect the kernel profile of betel nut, extract the image boundary, and obtain the

---

Y. Dai · M. Lu ( ) · Z. Chen

Hunan University of Science and Technology, Xiangtan 411201, China

e-mail: [mlu@hnust.edu.cn](mailto:mlu@hnust.edu.cn)

smooth kernel contour. Liu et al. [3] proposes the IMS-YOLO model to detect the tomato fruit of the greenhouse shed with an accuracy of 97.13%. At the same time, the detection speed has been improved. This method [4] provides a good solution to the problem of fruit overlap under natural conditions.

In this paper, the lightweight mobilenetv3 network model is used to identify and classify the kernel of betel nut in complex environments. Compared with other depth classification models, Mobilenetv3 has good parameter optimization characteristics. It can achieve fast and accurate classification effect for the key process to provide pre-work preparation.

## 2 A Method of Quick and Accurate Classification

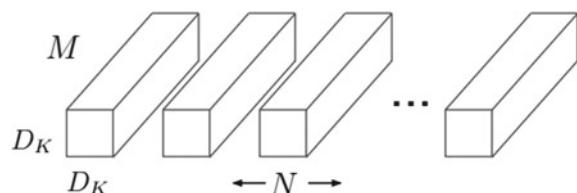
With the continuous development of deep network, it provides a powerful platform for image processing. At present, most deep network frameworks have the advantage of high accuracy. However, although the continuous improvement of trunk extraction network depth can bring high precision, but also will bring some column problems. The requirements of computing power gradually increase and the increase of parameters also brings a series of problems to be optimized to the classification model.

### 2.1 The Structure of Mobilenetv3

The introduction of the MobileNet series model is an important symbol of lightweight network applications and has changed the traditional convolutional approach [5]. The core idea of Mobilenetv3 is to propose a separable deep convolution structure, which use depthwise convolutions for compression, and use pointwise convolutions to reduce computation and speed up computing.

**Depthwise separable convolution:** Depthwise separable convolution consists of depthwise convolution and pointwise convolution. The series model changes the feature extraction method by depthwise separable convolution, which is greatly improved in the calculation parameters compared with the standard convolution, and improves the recognition speed. The structure of which is as follows (Fig. 1).

**Fig. 1** Standard convolution



When you enter a feature map of  $DF \times DF \times N$ , the standard convolution layer produces a feature map of the same size.  $DF$  is the value of the length and width of the input,  $M$  is the number of channels; Set the final generated output size to  $DG \times DG \times N$  feature map.  $DG$  is the value of the length and width of the output picture, and  $N$  is the number of output picture channels (output depth). The standard convolution layer calculates and parameterizes the feature map through the convolutional  $DK \times DK \times M \times N$ , Where  $DK$  is the spatial dimension of the kernel assumed to be square,  $M$  is the number of input channels, and  $N$  is the number of output channels as defined before [6].

Standard convolutions have the computational cost of:

$$C_S = D_K \times D_K \times M \times N \times D_F \times D_F \quad (1)$$

The calculation cost depends on the number of input channels, the number of output channels, the convolution size, and the size of the original feature map. MobileNets can be a good way to separate the structure and connections between them. First, it uses depth separate convolution to break the connection between the number of output channels and convolution.

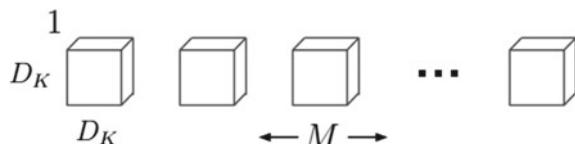
The standard convolution has the effect of filtering features and combining features [7]. The filtering and combination steps can be split into two steps, which via the use of depthwise separable convolutions in order to reduce computational costs Significantly (Fig. 2).

Depthwise separable convolution consists of two layers: depthwise convolutions and pointwise convolutions. We use depthwise convolutions to apply a single filter per each input channel. Pointwise convolution creates a linear combination of the output of the depthwise layer. MobileNets use both batchnorm and nonlinear activation function for both layers [8].

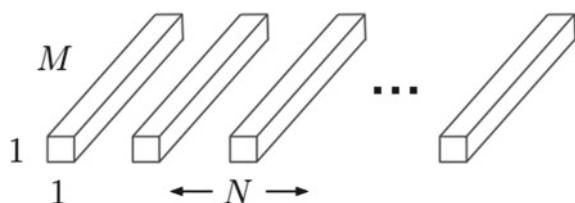
Depthwise convolution has a computational cost of (Fig. 3):

$$C_{DW} = D_K \times D_K \times M \times D_F \times D_F \quad (2)$$

**Fig. 2** Depthwise convolutions



**Fig. 3** Pointwise convolution



Depthwise convolution only filters input channels, it does not combine them to create new features. Therefore, an additional layer that computes a linear combination of the output of depthwise convolution by the use of  $1 \times 1$  convolution in order to generate new features.

Pointwise convolutions cost:

$$C_{PW} = D_K \times D_K \times M \times N \quad (3)$$

Therefore, the total calculated amount of the depthwise separable convolutions is:

$$C_{DS} = C_{DW} + C_{PW} \quad (4)$$

$$C_{DS} = D_K \times D_K \times M \times D_F \times D_F + D_K \times D_K \times M \times N \quad (5)$$

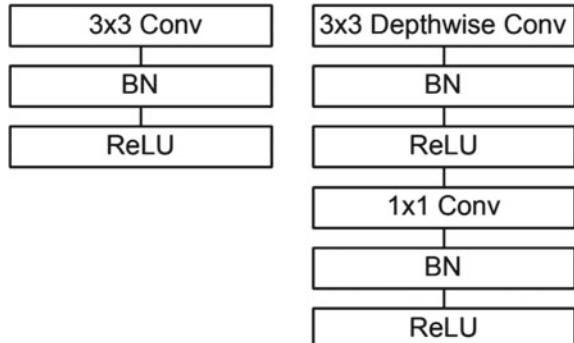
which is the sum of the depthwise and  $1 \times 1$  pointwise convolution.

By expressing convolution as two steps process of filtering and combining we get a reduction in computation of:

$$\begin{aligned} \frac{C_{DS}}{C_S} &= \frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} \\ &= \frac{1}{N} + \frac{1}{D_K^2} \end{aligned} \quad (6)$$

In conclusion, mobileNets uses  $3 \times 3$  depthwise separable convolutions which uses between 8 and 9 times less computation than standard convolutions [9] (Fig. 4).

**Fig. 4** Left: standard convolutional layer with BN and ReLU. Right: depthwise separable convolutions with depthwise and pointwise layers followed by BN and ReLU



### 2.1.1 Body Architecture of MobileNetv3

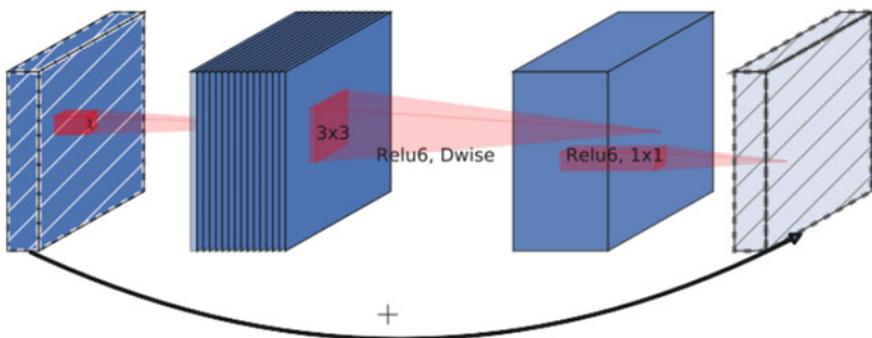
Compared with Mobilenetv2, the Mobilenetv3 model is an improvement based on the backbone extraction network of it. In the Mobilenetv3 model, a lightweight attention model based on the Squeeze and excitation structure (SE) was introduced, and the excitation function was improved (Figs. 5 and 6).

In Mobilenetv3, the SE module pools each channel of the input feature matrix and then obtains a one-dimensional vector through two fully connected layers. Its role is to obtain the weight relationship of the original feature matrix channel. Give high weight to important channels to improve feature extraction [10].

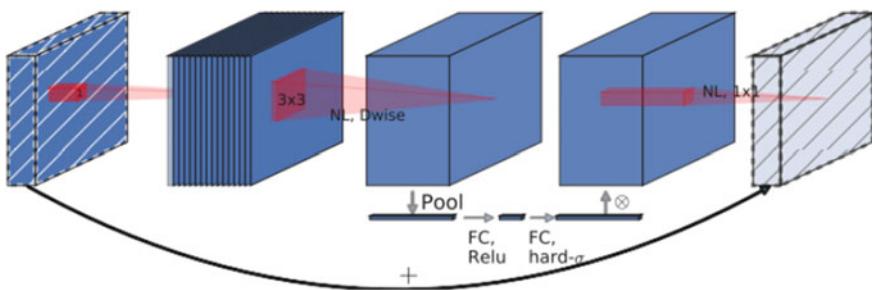
Mobilenetv3 abandoned the previous swish function and used h-swish function, as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (7)$$

$$\text{swish } x = x \cdot \sigma(x) \quad (8)$$



**Fig. 5** Block of Mobilenetv2



**Fig. 6** Block of Mobilenetv3

$$\text{ReLU6}(x) = \min(\max(x, 0), 6) \quad (9)$$

$$h - \text{swish}[x] = x \frac{\text{ReLU6}(x + 3)}{6} \quad (10)$$

According to condition (10), Compared to the ReLU and swish used by Mobilenetv1 and Mobilenetv2, the h-swish function is better nonlinear, easy to guide, and more quantified [11], which is also an advantage of Mobilenetv3. The main structure of Mobilenetv3 is shown in Fig. 7.

Input	Operator	exp size	#out	SE	NL	$s$
$224^2 \times 3$	conv2d	-	16	-	HS	2
$112^2 \times 16$	bneck, 3x3	16	16	-	RE	1
$112^2 \times 16$	bneck, 3x3	64	24	-	RE	2
$56^2 \times 24$	bneck, 3x3	72	24	-	RE	1
$56^2 \times 24$	bneck, 5x5	72	40	✓	RE	2
$28^2 \times 40$	bneck, 5x5	120	40	✓	RE	1
$28^2 \times 40$	bneck, 5x5	120	40	✓	RE	1
$28^2 \times 40$	bneck, 3x3	240	80	-	HS	2
$14^2 \times 80$	bneck, 3x3	200	80	-	HS	1
$14^2 \times 80$	bneck, 3x3	184	80	-	HS	1
$14^2 \times 80$	bneck, 3x3	184	80	-	HS	1
$14^2 \times 80$	bneck, 3x3	480	112	✓	HS	1
$14^2 \times 112$	bneck, 3x3	672	112	✓	HS	1
$14^2 \times 112$	bneck, 5x5	672	160	✓	HS	2
$7^2 \times 160$	bneck, 5x5	960	160	✓	HS	1
$7^2 \times 160$	bneck, 5x5	960	160	✓	HS	1
$7^2 \times 160$	conv2d, 1x1	-	960	-	HS	1
$7^2 \times 960$	pool, 7x7	-	-	-	-	1
$1^2 \times 960$	conv2d 1x1, NBN	-	1280	-	HS	1
$1^2 \times 1280$	conv2d 1x1, NBN	-	k	-	-	1

**Fig. 7** Body architecture of MobileNetv3

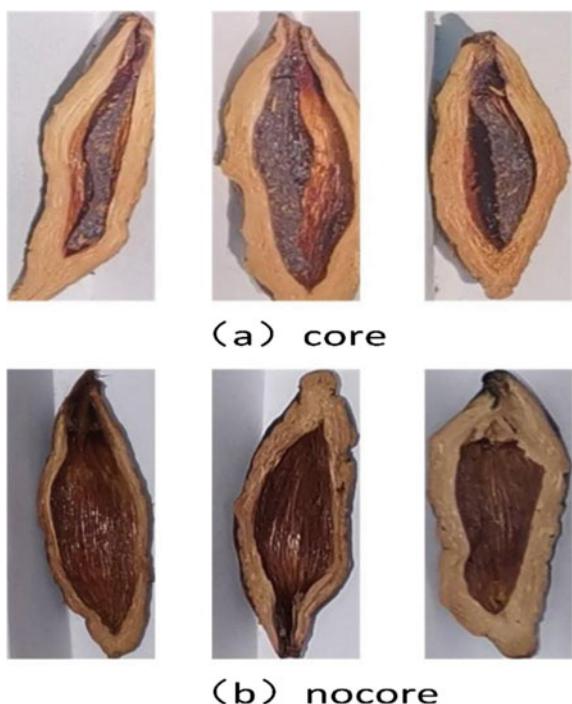
### 3 Analysis of Experimental Results

#### 3.1 Dataset Description

In this work, we first need to get a certain number of datasets to prepare for deep learning model training. The dataset is collected at the local betel nut processing plant and has a 2D RGB image resolution size. The size of the input image is almost  $300 \times 180$ . The betel nut fruit is less affected by the natural environment in the real situation, mainly considering the effect of light intensity on the recognition. Because the core target of the fruit is small and the background is difficult to extract, the betel nut image needs to be pre-treated before the experiment can be carried out.

To prevent under-diversity datasets from overfitting training models, we collected long-distance data and close-range data, which contains dark and bright data [12]. Totally 1200 images. In order to increase the diversity of data, this paper also adopts the mosaic enhancement method. This method randomly crops, flips, and stitches the dataset. Increasing the diversity of data is conducive to network training and classification. We packaged the pictures into two folders and then import them into the training model. We label images and divide the dataset into training and validation sets with the ratio of 8:2. The sample images in Fig. 8 are from the above said datasets

**Fig. 8** Core and no core samples



**Table 1** Training and testing ratio of betel nut images

Dataset	Total instance	Dark data	Bright data	Close-range data	Long-distance data
Training	975	467	500	430	537
Testing	245	113	120	86	167
Total	1200	580	620	516	704

and the number of betel nut objects used for training and testing is described in Table 1.

### 3.2 Evaluation Metrics

In order to evaluate the Mobilenetv3 model, the following parameters are used in this paper: Precision, Accuracy, Recall. The label of betel nut is classified as core and nocore. If the original sample label is nuclear and the model validation result label is the same, the sample is called True positive (TP). Similarly, the correct core validation sample is True Negative (TN) [13]. If the category classification errors are False Positive (FP) or False Negative (FN), condition (11)–(13) are utilized to measures the exactness, accuracy, and review of the item identification model.

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

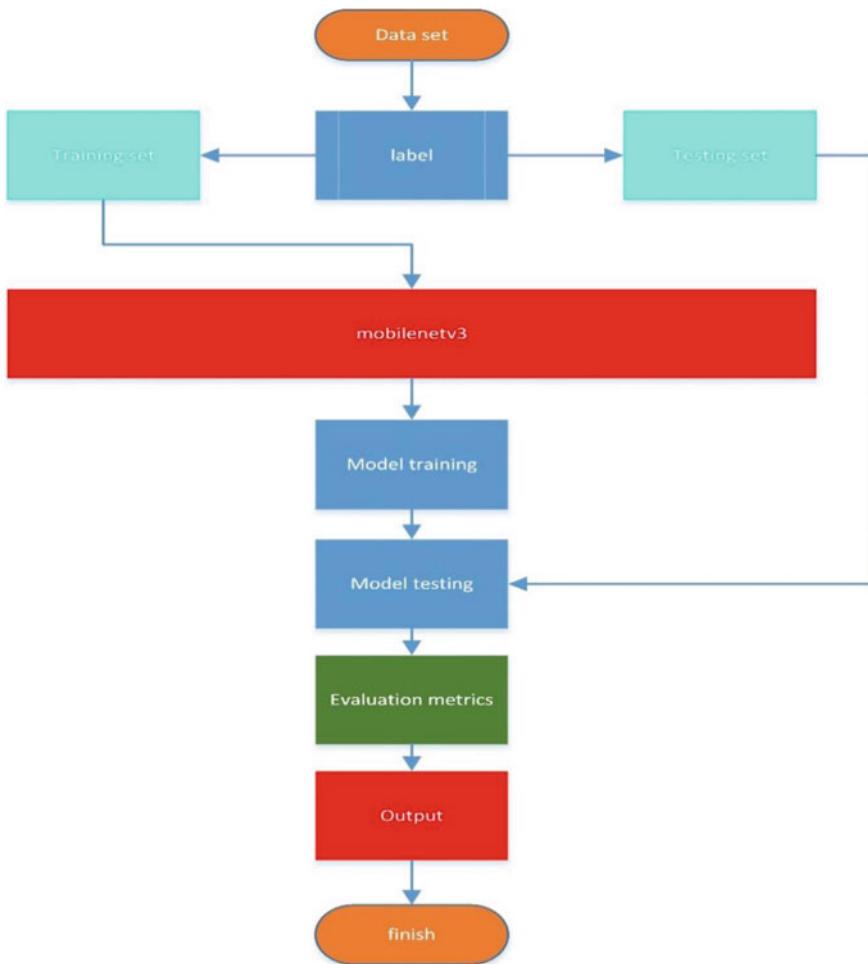
$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (13)$$

### 3.3 Model Training

To evaluate the classification effect of the dataset in the trained Mobilenetv3 model, All the experiments were trained and performed using PC with Intel i7 processors @ 3.40 GHz speed and 16 GB of RAM. This article uses Mobilenetv3 as an image feature extraction network. The picture of the dataset is preprocessed to enter the picture of the training set into the model of Mobilenetv3 for a training iteration [14]. We trained the network for 10 k epochs with learning rate of 0.0001 and Adam optimizers, and save trained models. This lab flowchart is shown in Fig. 9.

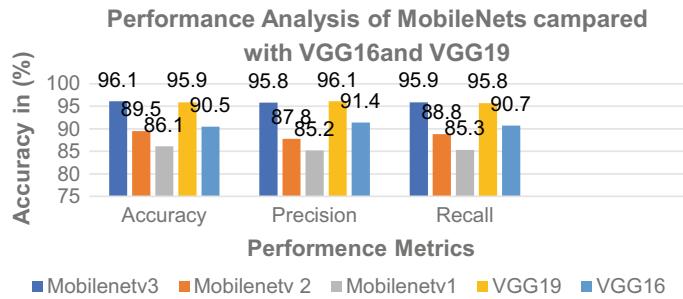
In this paper, Binary Cross entropy is used as a loss function. Use the softmax regression function to classify each picture and output confidence, condition (14).



**Fig. 9** Experimental flowchart

$$H_P(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (14)$$

This article experimented with Mobilenetv1 and Mobilenetv2. The advantages of Mobilenetv3 are demonstrated by comparative experiments with three models [15]. The results of the experiment are shown in Fig. 10. Mobilenetv1 and Mobilenetv2 have accuracy of 86.1% and 89.5%, respectively. However, the accuracy of the Mobilenetv3 model was 96.1%; Mobilenetv3 is also superior to Mobilenetv1 and Mobilenetv2 with 95.8% accuracy. Recall rate are also the best of the three models in Table 2.



**Fig. 10** Comparison analysis of MobileNets with VGG16 and VGG19

**Table 2** Performance comparison of Mobilenetv3 with Mobilenetv1 and Mobilenetv2

Model	Precision	Accuracy	Recall
Mobilenetv1	85.2	86.1	85.3
Mobilenetv2	87.9	89.5	88.8
Mobilenetv3	95.8	96.1	95.9
VGG19	96.1	95.9	95.8
VGG16	91.4	90.5	90.7

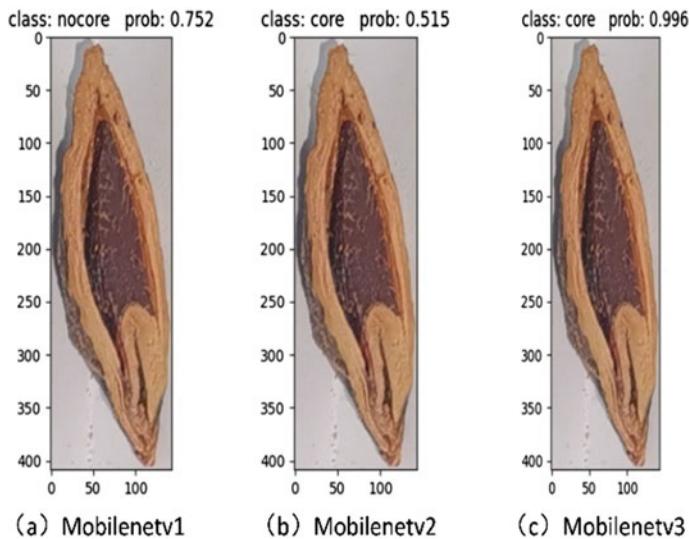
At the same time, Table 2 shows that in the deep network models of VGG16 and VGG19, the results of the classification indicators for the betel nut core are not always superior to Mobilenetv3. Mobilenetv3 not only gains an advantage in each evaluation metrics, but also has a certain degree in parameter optimization according to condition (6).

In the validation set sample, a sample is selected randomly and the evaluation indicators of the three models are compared in turn. Figure 11 shows the result of experiment.

It is clear in Fig. 11 that the sample was not properly classified in Mobilenetv1, and there was a degree of improvement in Mobilenetv2 and Mobilenetv3. Compared to the previous two models, mobilenetv3 has a significant improvement in accuracy, precision, and recall. The results show that, it can be classified correctly, but the confidence level of the label is only 0.515, far less than 1 in the Mobilenetv2 model. Compared to Mobilenetv2, Mobilenetv3 is not only classified correctly, but has a confidence level of almost 1. As a result, in Mobilenetv3, the evaluation metrics of image classification has been significantly improved.

## 4 Conclusion

Because the characteristics and background of the betel nut are difficult to distinguish, it is not possible to solve the classification problem perfectly. with the continuous



**Fig. 11** The results of comparative experiments on different models

development of deep learning, there are still many better ways to solve such problems. In this paper, a lightweight network model is used for the classification of the betel nut kernel. The model detects quickly with high accuracy compared to other depth models. It can be seen by experimental comparison, Mobilenetv3 performs the best in image classification and has the best evaluation metrics. At present, the labor costs of the betel nut processing are high. This article combines deep learning with betel nut classification and the lightweight network model is used for the operation of betel nut processing classification. This is a good attempt to make a bold guess and pave the way for the introduction of machine vision in later processing plants.

**Acknowledgements** This research was funded by the National Natural Science Foundation of China (grant number 61672226, 61903137). This research was funded by the Natural Science Foundation of Hunan Province (grant number 2020JJ4316).

## References

1. Yue, X., Jian, C., Ying, G.: Study on grading of betel nuts by computer vision. *Food Mach.* **32**(08), 91–94 (2016)
2. Liu, Z., Dong, Z., Ying, Z.: Betel nut stones contour detection based on semantic segmentation. *Comput. Technol. Autom.* **38**(04), 105–112 (2019)
3. Fang, L., Yu, L., Sen, L.: Fast recognition method for tomatoes under complex environments based on improved YOLO. *Trans. Chin. Soc. Agric. Mach.* **51**(06), 229–237 (2020)
4. Liang, H., Jian, C., Xian, X.: Research on recognition method for automatic orientating betel nut. *Food Mach.* **36**(12), 95–98 (2020)

5. Xia, Z., Ning, L., Rui, Z.: An improved lightweight network MobileNetV3 Based YOLOv3 for pedestrian detection. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 114–118 (2021)
6. Han, W.: A YOLOV3 system for garbage detection based on MobileNetV3\_Lite as backbone. In: 2021 International Conference on Electronics, Circuits and Information Engineering (ECIE), pp. 254–258 (2021)
7. Seunghyeok, B., Seongju, L., Sungho, S.: Robust skin disease classification by distilling deep neural network ensemble for the mobile diagnosis of Herpes zoster. *IEEE Access* **9**, 20156–20169 (2021)
8. Hai, P., Zai, P., Yao, W., Chen, L.: A new image recognition and classification method combining transfer learning algorithm and MobileNet model for welding defects. *IEEE Access* **8**, 119951–119960 (2020)
9. Yan, W., Jing, Y., Qi, S., Jun, L.: A MobileNets convolutional neural network for GIS partial discharge pattern recognition in the ubiquitous power Internet of Things context: optimization, comparison, and application. *IEEE Access* **7**, 150226–150236 (2019)
10. Suhas, S., Hardik, J., Olaf, H.: A power efficiency enhancements of a multi-bit accelerator for memory prohibitive deep neural networks. *IEEE Open J. Circ. Syst.* **2**, 156–169 (2021). <https://doi.org/10.1109/OJCAS.2020.3047225>
11. Sheng, B., Ying, Z., Min, D.: An embedded inference framework for convolutional neural network applications. *IEEE Access* **7**, 171084–171094 (2019)
12. Ling, G., Lin, Z., Zhao, W.: Hierarchical attention-based astronaut gesture recognition: a dataset and CNN model. *IEEE Access* **8**, 68787–68798 (2020)
13. Yan, S., Bo, P., Yi, F.: Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery. *IEEE Robot. Autom. Lett.* **6**(2), 3870–3877 (2021)
14. Metin, A., Yong, D.: Deep learning classification of systemic Sclerosis skin using the MobileNetV2 model. *IEEE Open J. Eng. Med. Biol.* **2**, 104–110 (2021)
15. Hyoukjun, K., Michael, P., Angshuman, P.: Flexion: a quantitative metric for flexibility in DNN accelerators. *IEEE Comput. Archit. Lett.* **20**(1), 1–4 (2021)

# PRM: Pose Recalibration Module for Action Recognition



Guixiong Tian<sup>ID</sup>, Yang Yi<sup>ID</sup>, Zijian Meng<sup>ID</sup>, Zhonghong Li<sup>ID</sup>,  
and Jialun Song<sup>ID</sup>

**Abstract** Two-stream convolutional network is the mainstream method of human action recognition, which can achieve excellent recognition precision in most datasets. However, the challenge of two-stream convolutional network is that it can not model human action well. To tackle this issue, this paper proposes Pose Recalibration Module (PRM) to better model human action features. The proposed PRM is composed of three components: (1) Part Affinity Fields pose estimator Module to explicitly capture human actions and generate human joint heatmap, (2) Multi-scale Action Extraction Module to construct different scales of joint heatmap sequence at different time intervals, and (3) Action Classification Module to classify action. Finally, this paper arranges late fusion strategy to fuse pose modalities and the two-stream network to obtain the final classification score. Experimental results on UCF101 and HMDB51 show that our approach can boost the performance of two-stream network. Meanwhile, our method obtains a competitive performance compared with state-of-the-art methods.

**Keywords** Pose modality · Feature fusion · Action recognition

## 1 Introduction

Human action recognition (HAR) is a classic research topic in the field of computer vision. For a given human action video, the computational model can extract the features in the video and identify the action category [1, 2].

Due to the strong generalization ability of deep networks, it has higher precision after fusing with different types of features. Therefore another popular research option based on deep learning is fusing with other types of modality, in which the pose

---

G. Tian (✉) · Y. Yi · Z. Meng · Z. Li · J. Song

School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China

e-mail: [tiangx@mail2.sysu.edu.cn](mailto:tiangx@mail2.sysu.edu.cn)

Y. Yi

Guangzhou Xinhua University, Guangzhou, China

Guangdong Province Key Laboratory of Big Data Analysis and Processing, Guangzhou, China

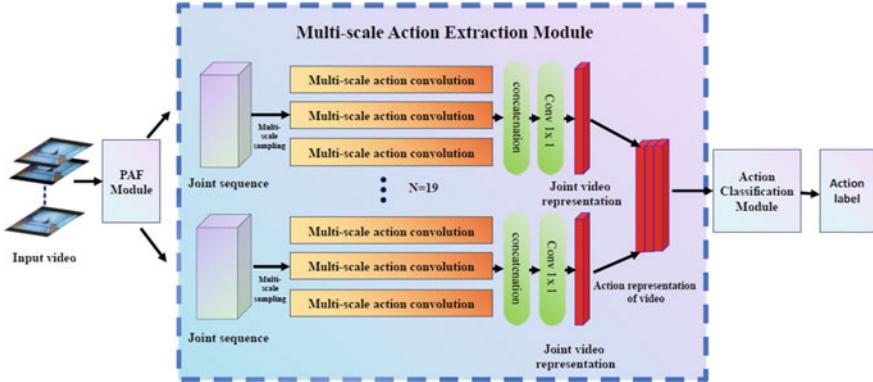
modality is a popular fusion mode. With the development of the field of pose estimation, the latest pose estimators, such as PAF [3], have high recognition precision, and have the characteristics of multi-target, real-time and end-to-end. Thus there is no need for an additional preprocessing to take the human pose as the input and it can be obtained by embedding the pose estimator. The pose modality includes trunk and facial features, which contains rich action information. However, the pose modality only contains position information and does not have pixel information, the expressive power of human action features is limited. Therefore, a better method is using human action features as the supplement of video pixel features to jointly determine the action label. Inspired by the above work, this paper proposes the Pose Recalibration Module (PRM), which consists of three submodules, Part Affinity Fields pose estimate Module (PAFM), Multi-scale Action Extraction Module (MAEM), and Action Classification Module (ACM). PAFM is used to generate the human pose modality. Then MAEM constructs a multi-scale action model by convolution human pose modality. It can obtain the video representation based on human actions. Finally, input the video representation into ACM to get the corresponding action label. Experimental results show that our method can achieve competitive performance compared with the state-of-the-art methods on two datasets UCF101 and HMDB51.

Overall, our contributions can be summarized as follows:

- A light-weight yet effective module PRM with PAFM, MAEM, and ACM, is proposed to recalibrate action features and activation.
- We conduct extensive experiments on two human action recognition datasets UCF101 and HMDB51, and compare PRM-based methods with the state of the arts to verify the performance of our module.

## 2 Approach

To make use of human action information efficiently, this paper proposes a Pose Recalibration Module (PRM). As Fig. 1 shown, the overall architecture of our PRM can be divided into three submodules: PAFM, MAEM, and ACM. The end-to-end PAFM is used to obtain the human pose modality of the video. The modality is the heatmap of each joint position of the human body in the video frame. Then the action features are extracted from these heatmaps on the spatial-temporal scale by MAEM, and the action-based video representation is obtained by fusing the features of each joint. Finally, ACM identifies the action according to the video representation and gets the corresponding action label. The following subsections describe the details of each submodule.



**Fig. 1** The overview of pose recalibration module

## 2.1 PAFM

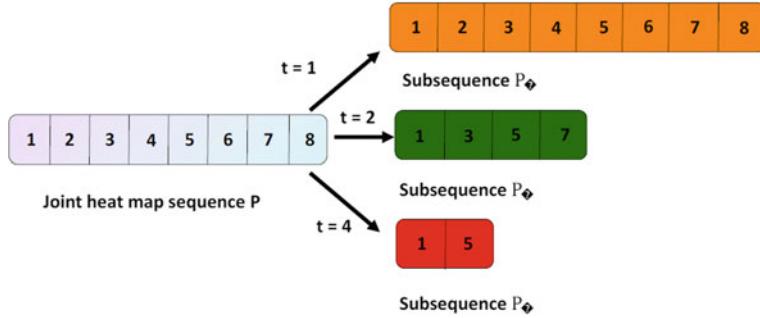
To obtain motion of the human body explicitly, this paper constructs a pose estimation network based on PAF to generate the human pose modality. For the input video, the convolution feature is generated by VGG. Then the feature is input into two branches to generate Part Confidence Maps (PCM) and PAF respectively. PCM is the heatmap of the confidence of each joint and PAF is the trunk confidence of connecting two joints, including position and orientation information. The two generated branches will be fused, and then input to the next branch to start a new round of iteration, a total of 7 iterations.

This paper mainly uses the joint heatmap of PCM after 7 iterations as the input of the next process. PCM contains  $N$  kinds of human joints. This paper use heatmap of 19 kinds of human joints. For input video  $S \in R^{T \times H \times W}$ ,  $T$  is length of video,  $H$  and  $W$  are the height and width of the video respectively. The sequence of joint thermogram extracted by PCM is  $P \in R^{N \times T \times H \times W}$ , Where  $N$  is the number of joints.

## 2.2 MAEM

After obtaining the heatmap of human joints, this paper proposes a multi-scale sampling method, which constructs different scale heatmap sequences at different time intervals. Then the action features are extracted by using group convolution for each sequence. Finally, the action-based video representation is obtained by fusing different scale sequences.

As shown in Fig. 2, for the input  $n$ -th joint heatmap sequence  $P_n \in R^{T \times H \times W}$  ( $n = 1, \dots, N$ ), different sampling intervals  $t$  are used for sampling. In this paper, the values of  $t$  are  $t_k = \{1, 2, 4\}$ . So we get a total of  $K = 3$  subsequences of joint heatmap  $P_{nk} \in R^{\frac{T}{t} \times H \times W}$ , the length of each subsequence is  $T/t$ . For the



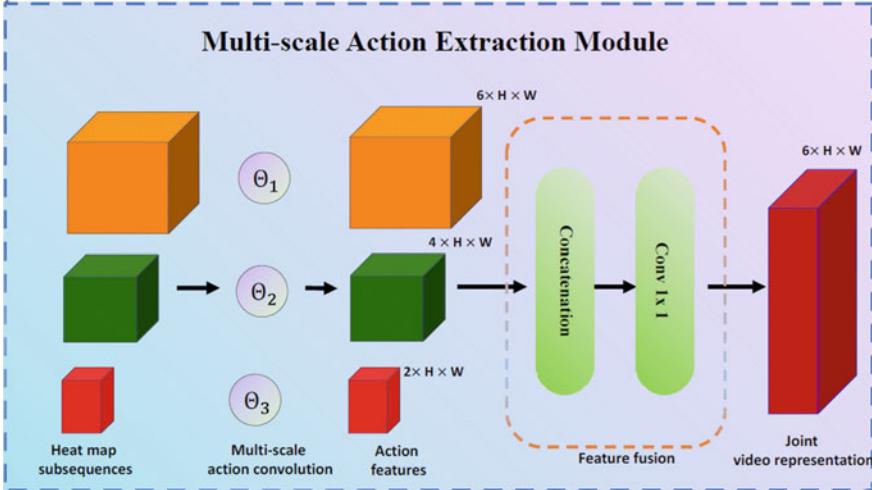
**Fig. 2** Multi-scale joint heatmap sequence sampling

sequence with small sampling interval, more samples are sampled, which contains rich semantic information and action information. It can be used as the main part of action feature extraction. For the sequence with large sampling interval, the number of samples is small and mainly contains the action information of different scales. It can be used as a supplementary part to extract the action features.

To give full play to the characteristics of these sequences, different sizes of convolution kernel is used to extract the action features of different sequences. It allocates more output channels for the sequences with small intervals and fewer output channels for the sequences with large intervals. At the same time, due to the sparse features and concentrated distribution of the heatmap, the use of large-scale convolution will increase the cost of parameters and not improve the accuracy. Hence small-scale convolution is more suitable. Based on the above analysis, MAEM is proposed to extract action features. MAEM uses  $1 \times 1$  2D convolution kernel with different number of output channels to extract different sequences. Thus the convolution kernel size of each subsequence is  $k \in R^{O_k \times T_k \times 1 \times 1}$ . The input and output channels of convolution kernel change with the length of subsequence. Different scale subsequences extract their own action features. Finally, through the concatenation and  $1 \times 1$  convolution compression feature is used to get the action feature of the joint. As shown in Fig. 3, the whole process can be summarized as follows:

$$\begin{cases} \widehat{P}_{nk} = P_{nk} * k, (k = 1, \dots, K). \\ \widehat{P}_n = conv_{1 \times 1} [\widehat{P}_{n1}, \dots, \widehat{P}_{nk}]. \end{cases} \quad (1)$$

where  $P_{nk}$  represents the subsequence sampled by the joint,  $\widehat{P}_{nk}$  is the action feature corresponding to the subsequence,  $k$  represents the action convolution corresponding to the subsequence,  $\widehat{P}_n \in R^{Q \times H \times W}$  is the video action representation of the joint sequence obtained by aggregating all the subsequence action features,  $Q$  is the output channel of joint action. In this paper, the value is 6, the value will be specified in the experimental section.



**Fig. 3** The overview of multi-scale action extraction module

In order to supplement the number of samples and prevent overfitting, the same scale sequence of different joints will share the weight of convolution kernel. The video representation of all joints is concatenated to get the final video representation. The process is shown in the following formula:

$$\hat{P} = [\hat{P}_1, \dots, \hat{P}_n] (n = 1, \dots, 19). \quad (2)$$

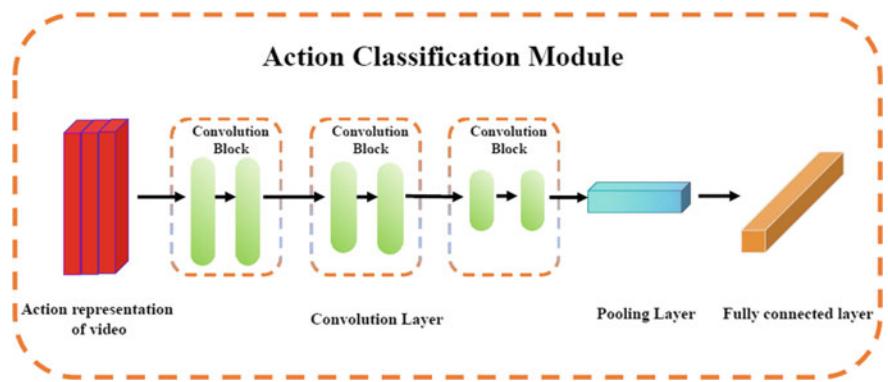
where  $n$  is the number of joints, this paper uses a total of 19 joints of the heatmap sequence, so the final generated action-based video representation  $\hat{P} \in R^{114 \times H \times W}$ , 114 channels in total. The convolution method in this paper can save parameters and extract the semantic information of the input heatmap. Through the construction of different scale sequences to extract a variety of action information, the video representation has more expressive power and robustness.

## 2.3 ACM

The video representation obtained by MAEM is input to ACM for classification. As the input of processing is middle-level semantic features, compared with the original image, the features are sparser. Thus it cannot use deep network training. Therefore, the structure design of action recognition network is referred to reference [4], and the structure is shown in Table 1 and Fig. 4. ACM consists of three convolution blocks, a global pooling layer and a fully connected layer. Each convolution block contains two convolution layers. The size of the convolution kernel is  $3 \times 3$ . The first convolution

**Table 1** ACM network structure

Network layer	Size of output	Parameter setting
Input	$114 \times H \times W$	
conv1_1	$128 \times H/2 \times W/2$	conv $3 \times 3$ , stride 2
conv1_1	$128 \times H/2 \times W/2$	conv $3 \times 3$ , stride 1
conv2_1	$256 \times H/4 \times W/4$	conv $3 \times 3$ , stride 2
conv2_2	$256 \times H/4 \times W/4$	conv $3 \times 3$ , stride 1
conv3_1	$512 \times H/8 \times W/8$	conv $3 \times 3$ , stride 2
conv3_2	$512 \times H/8 \times W/8$	conv $3 \times 3$ , stride 1
FC-512	$512 \times 1 \times 1$	AvgPool
Classification	$K \times 1 \times 1$	

**Fig. 4** The overview of ACM

kernel stride is 2, and the second convolution kernel stride is 1. Compared with the previous convolution block, each convolution block doubles the number of channels. Therefore, after a convolution block, the resolution of the feature will be compressed by half, and the number of channels will be doubled. After the convolution block, the features are compressed in the global pooling layer, and finally the action label is obtained after the fully connected layer.

### 3 Experiments

In this section, we first discuss the experiment setting. Then we conduct experiments on channel number of video representation to obtain the optimal sequence number of scale sampling in pose modality. And PRM proposed in this paper is compared with the same type of methods to verify the competitiveness of each module of

the proposed method. Finally, to validate the effectiveness of the proposed PRM, data experiments are performed on two popular datasets HMDB51 [5] and UCF101 [6]. The results display that our PRM is competitive with several state-of-the-art methods: iDT + FV [7], TDD + FV [8], Two-stream (VGG16) [9], C3D [10], T3D [11], TSN [12], PA3D [13], Hidden Two-stream (TSN) [14], Coarse-to-fine (Motion) [15], MRST-T [16], SGN [17], ISTA [8], ISPAN [18], AARM [19].

### 3.1 Experiments Setting

The experimental environment is ubuntu14.04, 2.1 GHz 32 core processor, four GTX 1080 Ti, 64G memory, and 1.5 T storage space. The input video frame size is cut to  $224 \times 224$ , horizontal flip, random clipping, and corner clipping are used for data augmentation.

The parameters of PRM network training are the same as the optical flow network in AARM [19], but the input is 8 video frames and 400 epochs are trained. The overall evaluation standard uses Top1 accuracy. Similar to the fusion of two stream structures, the method in this paper adopts the late fusion method for the PRM network and the AARM two stream convolution network. Specifically, the two modes are carried out independently in the training stage, and only in the test stage can they be fused. The fusion method is to sum the scores of each category of the input video obtained from the two modes by weighting. The values in this paper refer to the method of previous work [4, 13] to set the same weight.

Due to the sparsity of the human pose modality, it is difficult to get better recognition results only by using the pose modality. Therefore, the PRM is combined with AARM two-stream convolution network [19].

### 3.2 Experiment on Channel Number of Video Representation

This experiment is to study the effect of the number of video channels  $Q$  of each joint sequence output on PRM in the action modeling stage. The whole video represents the total number of  $N \times Q$  channels, and the fixed number of joints  $N$  is 19. Therefore,  $Q$  controls the number of channels of the whole video representation. Experiments will test  $Q$  with different values. The experimental results are shown in Table 2.

It can be seen that the best performance can be obtained when  $Q$  is 6. It can be inferred that the reason is that when the number of channels is insufficient, it is difficult for each joint to save multi-scale action information. When the number of channels is too many, the information is redundant, which increases the training difficulty of the action recognition network and reduces the accuracy. Therefore, based on the experimental results, it is recommended to use the configuration of  $Q = 6$  on the two data sets of this experiment.

**Table 2** The number of channels for video representation results

Value of $Q$	UCF101 (%)	HMDB51 (%)
$Q = 2$	58.2	43.7
$Q = 4$	60.2	46.8
$Q = 6$	60.9	47.2
$Q = 8$	60.6	47.1
$Q = 10$	60.4	46.8

**Table 3** Comparison of pose-based recognition methods on split 1 of UCF 101 and HMDB 51

Fusion strategy	UCF101	HMDB51
PoTion [4]	—	43.4%
PA3D [13]	—	46.7%
PRM (ours) (%)	<b>60.9</b>	<b>47.2</b>

### 3.3 Comparison with Pose Modules

This experiment focuses on the difference between PRM proposed in this paper and other pose-based action recognition methods, such as PoTion and PA3D. The input of the comparison module is the heatmap of each joint to achieve fair competition. The experimental results are shown in Table 3.

From the experimental results, it can be observed that the PRM proposed in this paper has the best recognition effect on HMDB51. PoTion uses the action modeling of merging heatmaps along with time series. It is not as good as the other two convolution-based action modeling methods. Compared with PA3D, this paper uses multi-scale sampling, which can lead by a small margin. It shows that the PRM has stronger expression ability in multi-scale motion modeling ability, and verifies the effectiveness of the PRM.

### 3.4 Comparison with the state of the arts

This paper finally compares our proposed PRM with various state-of-the-art action recognition methods. As shown in Table 4, compared with other methods using 2D convolution network, the recognition precision of this method is comprehensively leading, which shows that 2D convolution network still has a lot of room for improvement, and human action modality can play an effective role. On the other hand, the proposed method still achieves competitive results with some 3D convolution networks without using 3D convolutional network structure, which shows that the proposed method has good performance. Moreover, the proposed method achieves the best recognition rate after using the TSN structure, which verifies the effectiveness and competitiveness of the proposed human action recognition method based on pose feature fusion.

**Table 4** Comparison with the state of the arts on split 1 of UCF101 and HMDB51

Methods	UCF101	HMDB51
iDT + FV [7] (%)	85.9	57.2
TDD + FV [8] (%)	90.3	63.2
Two-stream (VGG16) [9] (%)	86.9	58.4
C3D [10]	85.2%	–
T3D [11] (%)	91.7	61.6
TSN [12] (%)	94.0	68.5
PA3D [13]	–	55.3%
Hidden two-stream (TSN) [14] (%)	93.2	66.8
Coarse-to-fine (motion) [15] (%)	93.6	69.3
MRST-T [16] (%)	92.2	68.9
SGN [17]	90.5%	–
ISTA [8] (%)	87.1	53.1
ISPAN [18] (%)	94.8	64.6
AARM [19] (%)	93.4	67.6
<b>PRM + AARM (ours) (%)</b>	<b>94.8</b>	<b>69.1</b>
<b>PRM + AARM + TSN (ours) (%)</b>	<b>96.7</b>	<b>71.3</b>

## 4 Conclusions

In this paper, PRM module is introduced and divided into three submodules to better extract human action features, which solves the defect that the deep learning method ignores the explicit modeling of human action features. Through extensive experiments on two video datasets UCF101 and HMDB51, the effectiveness of the proposed method is verified, and the optimal hyper-parameter configuration is given for the datasets. At the same time, compared with other similar methods and current representative research results, the experimental results show that the proposed method can achieve competitive classification results.

**Acknowledgements** This work is partly supported by Guangzhou Science and Technology Project with No. 202002030273 and No. 201804010265, National Natural Science Foundation of China (NSFC No. 61672546), also by Key Discipline Project 2020XZD02, Xinhua College of Sun Yat-sen University.

## References

1. Hutchinson, M., Vijay, G.: Video action understanding: a tutorial. arXiv preprint [arXiv:2010.06647](https://arxiv.org/abs/2010.06647) (2020)
2. Pareek, P., Ankit, T.: A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif. Intell. Rev. **54**(3), 2259–2322 (2021)

3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
4. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: Potion: Pose motion representation for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7024–7033 (2018)
5. Huang, H., Garrote, H., Poggio, E., Serre, T., Hmdb, T.: A large video database for human motion recognition. In Proceedings of the IEEE International Conference on Computer Vision, vol. 4(5), 6 (2011)
6. Soomro, K., Zamir, A.R., Shah, M.: A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision, vol. 2(11) (2012)
7. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
8. Meng, L., Zhao, B., Chang, B., Huang, G., Sun, W., Tung, F., Sigal, L.: Interpretable spatio-temporal attention for video action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, p. 0 (2019)
9. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv: p. 1406.2199 (2014)
10. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatio-temporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
11. Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Van Gool, L.: Temporal 3d convnets: new architecture and transfer learning for video classification. arXiv preprint arXiv: pp. 1711. 08200 (2017)
12. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36. Springer, Cham (2016)
13. Yan, A., Wang, Y., Li, Z., Qiao, Y.: PA3D: Pose-action 3D machine for video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7922–7931 (2019)
14. Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A.: Hidden two-stream convolutional networks for action recognition. In: Asian Conference on Computer Vision, pp. 363–378. Springer, Cham (2018)
15. Ji, Y., Zhan, Y., Yang, Y., Xu, X., Shen, F., Shen, H.T.: A Context knowledge map guided coarse-to-fine action recognition. IEEE Trans. Image Process. **29**, 2742–2752 (2019)
16. Wu, H., Liu, J., Zha, Z.J., Chen, Z., Sun, X.: Mutually reinforced spatio-temporal convolutional tube for human action recognition. In: IJCAI, pp. 968–974 (2019)
17. Yu, T., Wang, L., Da, C., Gu, H., Xiang, S., Pan, C.: Weakly semantic guided action recognition. IEEE Trans. Multim. **21**(10), 2504–2517 (2019)
18. Du, Y., Yuan, C., Li, B., Zhao, L., Li, Y., Hu, W.: Interaction-aware spatio-temporal pyramid attention networks for action classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 373–389 (2018)
19. Li, Z., et al.: AARM: Action attention recalibration module for action recognition. In: Asian Conference on Machine Learning. PMLR (2020)

# Stereo Visual SLAM System Reasonably Use Point and Line Features



Huiyue Qiao , Xuhu Ren , Luyan Niu , Yang Feng , and Songzho Liu

**Abstract** Traditional feature-based SLAM systems rely on point features in the environment to recover the camera pose and build an environmental map. With the in-depth research of scholars, in order to make up for the disadvantages of point features in a low texture environment, stereo visual SLAM systems that combine both points and line segments are proposed. Although the stereo visual SLAM systems that combine both point features and line features improve the accuracy of tracking and they also increase the computational burden of the computer and reduce the tracking efficiency. However, in the actual environment, the environment will not always be in a state of low texture, so our work considers that the line feature can be selectively used. We selectively use line features during the tracking process and use line features as a supplement to point features in a low-texture environment. The main content of our work is proposing an analysis method that analyzes the change of the environmental characteristics that have been tracked during the tracking process to make the SLAM system possible to make good use of point features and line features in the tracking process. We test our system on public datasets and compare our results with state-of-the-art methods. The test results show that our stereo visual SLAM method can obtain more accurate results than the stereo visual SLAM system that uses points and even the stereo visual SLAM system that combines both point features and line features.

**Keywords** Stereo visual SLAM · Point features · Line features · Tracking feature analytical method

## 1 Introduction

In recent years, the reliability of visual simultaneous localization and mapping (SLAM) has gradually increased, and it has been increasingly used in many fields

---

H. Qiao · X. Ren ( ) · L. Niu · Y. Feng · S. Liu

College of Oceanography and Space Informatics, China University of Petroleum (East China),  
Qingdao, China

e-mail: [rjh@upc.edu.cn](mailto:rjh@upc.edu.cn)

[1]. And with the improvement of the computing performance of computers, the functions of SLAM are also becoming more and more complete.

SLAM systems can be divided into topological (e.g. [2–5]) and metric approaches. In our work, we focus on metric approaches, which use physical information in the environment and build a map with meaningful physical information. These approaches can be further categorized into feature-based and direct methods.

The direct methods are under the assumption that the brightness is constant. They get pose information by minimizing the photometric errors between consecutive frames and do not need to know the positional relationship between points. Recent representative works in the use of direct methods are semi-direct methods (SVO) [6] and sparse direct methods (DSO) [7]. Direct methods have the advantage of direct input images without considering the intermediate process, so the processing speed is fast. However, direct methods are very sensitive to environment brightness changes and are restricted to small-scale motions. Their effect is not good when large-scale motions occur. In contrast, feature-based methods extract stable feature information between frames, which have a certain degree of light stability and rotation invariance [8, 9].

In recent years, one of the most famous feature-based visual SLAM systems is ORB\_SLAM [9], which has a stable front-end, tracking process, back-end and forms a relatively mature visual SLAM framework. At the same time, ORB\_SLAM has achieved real-time results. But ORB\_SLAM is only based on point features that do not perform well in low-texture environments. Therefore, some scholars have proposed the use of more stable features in the environment, such as line features.

The visual SLAM systems that combine both point features and line features are proposed by scholars to obtain higher stability during tracking. But these methods will inevitably increase the burden on the computer and reduce the efficiency of tracking. However, in the actual environment with enough features, accurate pose information can be obtained without the participation of line features. Using line features at this time will add additional computational burden and can't help improve the tracking accuracy. And in a high-texture environment, too many line features will make the line features lose the representativeness of the environment and lead to a decrease in tracking accuracy. Therefore, our work uses line features as a method to complement point features during the tracking process. We design a method to determine when line features are needed for point feature supplementation based on experimental analysis. Our system is tested on different public datasets. The experimental results prove that our system can avoid use dense line features and obtains higher tracking efficiency and accuracy compared with the stereo visual SLAM system that combines both point features and line features. In summary, our contributions in this work are as follows:

- Propose a SLAM system that can reasonably use point features and line features. Our system can select the use of features according to the environmental information obtained by the system. When the point features are sufficient, only the point feature is tracked, and when the point features are not sufficient, the line features are used to supplement the point features.

- Propose a Point or Point and Line Method for judging the change of feature sparsity and effectiveness in the environment. This method can make stereo visual SLAM be able to predict environmental information according to the changes in feature state during the tracking process and decide whether line features are needed for point features supplementation.

Our paper is organized as follows: Section 2: Related work, Section 3: SLAM system overview, Section 4: Point or Point and Line Method, Section 5: Experimental Validation, Section 7: Conclusion.

## 2 Relate Work

The traditional feature-based SLAM tracks the key points of consecutive frames to obtain the relationship between the key points between consecutive frames and then minimize re-projection error functions to estimate the pose of the robot [1]. The current representative methods are Fast-SLAM [10], PTAM [11], and the most recent ORB-SLAM [9]. Fast-SLAM is a filter-based SLAM system that uses a filter-based method to optimize the pose after tracking to a point in the environment. PTAM is a monocular, key-frame-based SLAM system, which first proposes the system structure that splits camera tracking and mapping into parallel threads. ORB-SLAM uses fast and stable ORB features [12] for feature description in the environment. It extends PTAM and proposes a three-thread SLAM system that integrates tracking, local mapping, and loop closure. The three-threaded framework combines with map information is one of the most stable frameworks for feature-based SLAM systems. But in a low-texture environment, the point features tracking methods will cause tracking failure and reduce accuracy due to insufficient surrounding environment features. Therefore, some scholars consider using other spatial features in the environment to solve the adverse effects of point feature tracking in a low-texture environment, such as edge features, plane features, or line features.

In our work, we mainly consider line features. J. Neira proposed the first monocular vision SLAM system using vertical line segments [13]. As the optimized framework based on keyframes has become the mainstream of the visual SLAM system, GeorgKlein proposed a SLAM system that uses point features and line features based on keyframe optimization [14]. The key frame-based SLAM system is currently the most complete and stable framework achieved in ORB\_SLAM [9]. Therefore, Albert Pumarola [15] integrated line featured into the ORB\_SLAM framework and proposed a monocular vision SLAM system that combines LSD line features [16] with point features. This system detects line segments through endpoints and then performs descriptor-based tracking. Then Ruben Gomez-Ojeda combined the semi-direct method and the line feature to propose a stereo PL-SLAM [17], which is different from the monocular PL-SLAM. Firstly, it is based on a stereo system. Secondly, Stereo PL-SLAM does not make any assumption regarding the position of

the line's endpoints, but the STVO used by it has the problem of low stability. Therefore, Kun et al. [18] proposed ORB\_Line\_SLAM, which combines ORB\_SLAM stable front-end with monocular PL-SLAM line features process method. This system expands the ORB\_SLAM framework and obtains better results than stereo PL-SLAM in the public dataset test.

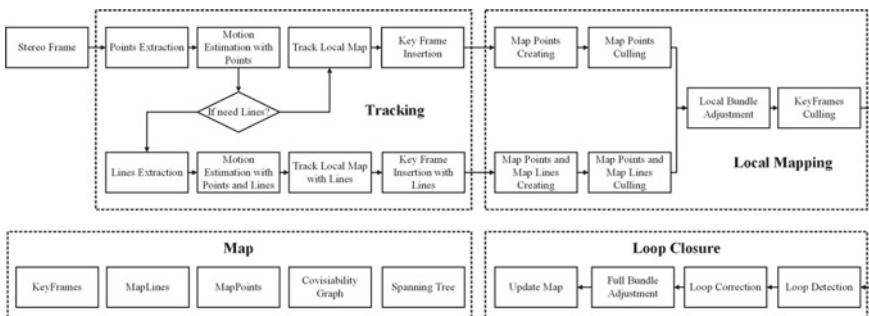
The above methods use point features and line features throughout the tracking process. However, in the actual environment tracking, there is no need to use the line feature all the time. When the environment texture is sufficient, the use of the line feature will only increase the burden on the computer and reduce the efficiency of tracking. At the same time, when in a high-texture environment, the introduction of too many line features will reduce the tracking accuracy. Therefore, how to judge whether to use line features is the main research content of our work.

### 3 SLAM System Overview

Our paper designs a SLAM system that can reasonably use point features and line features according to the changes in the feature state during the tracking process. We incorporate a method that can judge the effectiveness of point features during the tracking process to use point features and line features reasonably. The system is shown in Fig. 1, which is similar to ORB\_SLAM has four parts: Tracking, Map, Local Mapping, and Loop Closure.

#### 3.1 Tracking

This thread is to track the motion process between frames. We first use the method in [9] to extract the ORB features from the input pictures, then perform motion model tracking and pose-only optimization. After the motion model tracking, the



**Fig. 1** SLAM system overview

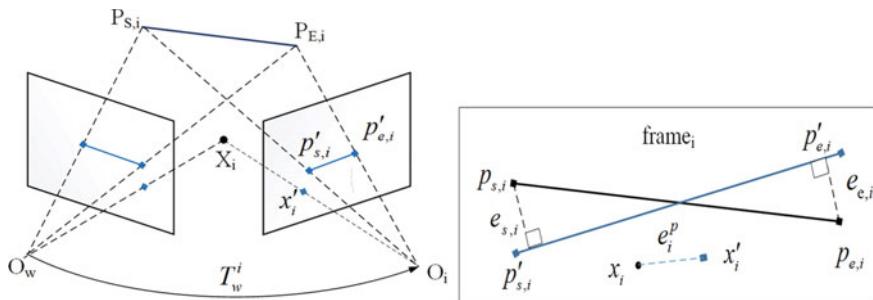
effectiveness of point features is judged. When our method judges the point feature tracking is sufficient, the processing of the next frame will continue. When our method judges the point feature tracking is insufficient, the current part of the environment will be tracked using the point features and the line features. The judgment method will be explained in detail in the fourth part. At the end of this thread, the current frame will be judged whether to insert a frame as a keyframe by the method in ORB\_SLAM [9].

### 3.2 Map

This part saves the map information during the tracking process, including the pose information and label of the keyframes. It also includes spatial 3D point features in point feature tracking mode, spatial 3D point features and spatial 3D line features in point feature and line feature tracking mode, visibility graph which contains visibility information of keyframes [19] and spanning tree.

### 3.3 Local Mapping

This process starts when a new keyframe is inserted. When a connection is detected in several local frames, several frames of pose information and its associated feature information will be optimized. The main part of this process is the construction of the loss function for bundle adjustment with points and lines. In Fig. 2,  $X_i \in \mathbb{R}^3$  be the 3D point of the point feature observed by the  $i$ -th keyframe,  $x_i \in \mathbb{R}^2$  be the image coordinates,  $T_w^i \in SE(3)$  be the pose of the  $i$ -th keyframe. According to the pose and 3D point, the re-projection can be expressed as  $\pi(T_w^i, X_i)$ . The re-projection error model of the spatial point feature can be expressed as follow:



**Fig. 2** The re-projection error of line segment feature and point feature

$$\mathbf{e}_i^p = \left[ \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \pi(\mathbf{T}_w^i, \mathbf{X}_i) \right] \quad (1)$$

The re-projection error of the line feature is constructed in the form of [15], In Fig. 1,  $\mathbf{P}_{S,i}, \mathbf{P}_{E,i} \in \mathbb{R}^3$  be the endpoints of line segments in space,  $\mathbf{p}'_{s,i}, \mathbf{p}'_{e,i} \in \mathbb{R}^2$  be image coordinates of the endpoints of the line segments,  ${}^h\mathbf{p}_{s,i}, {}^h\mathbf{p}_{e,i} \in \mathbb{R}^3$  are their corresponding homogeneous coordinates. The unit normal vector line of the plane is thus formulated by [15]:

$$\mathbf{l}_i = [l_0 \ l_1 \ l_2]^T = \frac{{}^h\mathbf{p}_{s,i} \times {}^h\mathbf{q}_{e,i}}{|{}^h\mathbf{p}_{s,i} \times {}^h\mathbf{q}_{e,i}|} \quad (2)$$

Combining the pose information, the two endpoints of the line feature are re-projected homogeneous coordinates are expressed as  ${}^h\pi(\mathbf{T}_w^i, \mathbf{P}_{S,i}), {}^h\pi(\mathbf{T}_w^i, \mathbf{P}_{E,i})$ , let  $\mathbf{e}_{s,i}, \mathbf{e}_{e,i}$  be the corresponding to the error of the start and end of the line segment respectively, the formula is as follows:

$$\mathbf{e}_i^l = \begin{bmatrix} \mathbf{e}_{s,i} \\ \mathbf{e}_{e,i} \end{bmatrix} = \begin{bmatrix} \mathbf{l}_i \cdot {}^h\pi(\mathbf{T}_w^i, \mathbf{P}_{S,i}) \\ \mathbf{l}_i \cdot {}^h\pi(\mathbf{T}_w^i, \mathbf{P}_{E,i}) \end{bmatrix} \quad (3)$$

When tracking point features, K is the set of keyframes, and P is the set of key points contained in the keyframes,  $\Omega_{\mathbf{e}_i^p}$  be the information matrix of the re-projection error of points,  $\rho$  be the robust kernel function. Then, the loss function can be formulated as:

$$C_p = \sum_{i \in K} \rho \left[ \sum_{j \in P} (\mathbf{e}_i^p)^T \Omega_{\mathbf{e}_i^p}^{-1} \mathbf{e}_i^p \right] \quad (4)$$

When the tracking point and line features,  $\Omega_{\mathbf{e}_i^l}$  be the information matrix of the re-projection error of lines, Then, the loss function including point and line features can be formulated as [15]:

$$C_{p,l} = \sum_{i \in K} \rho \left[ \sum_{j \in P} (\mathbf{e}_i^p)^T \Omega_{\mathbf{e}_i^p}^{-1} \mathbf{e}_i^p + \sum_{k \in L} (\mathbf{e}_i^l)^T \Omega_{\mathbf{e}_i^l}^{-1} \mathbf{e}_i^l \right] \quad (5)$$

According to the partial derivative of the loss function, the Jacobian matrix required for Gauss–Newton iterative optimization can be obtained. After executing the local bundle adjustment, the keyframes in the map information will be filtered.

### 3.4 Loop Closure

The loop closure process is to deals with error accumulation and performs global bundle adjustment optimization when a close loop constraint is established. This process uses the DBoW2 model [20] to search and measure the similarity between keyframes. In ORB\_SLAM, the author proved its excellent loop closure effect under normal conditions, so the loop closure process adapts the same idea as ORB\_SLAM [9].

## 4 Point or Point and Line Method

In the process of test in different public datasets, we found that when the environment is about to have insufficient effective features, it will reflect a certain degree of trend and regularity before it appears. Our method finds this trend and regularity by processing the original data and uses line features to supplement only in a time when the environmental features are insufficient. The specific implementation is as follows:

Firstly, we find the local transformation trend of the number of point features that are effectively matched after optimization can reflect the changing trend of the local environment to a certain extent.  $Y_i (i = 1, \dots, 10)$  is the number of optimized point features in the last ten frames. Perform a linear regression on it to get the trend flag  $\hat{\beta}_1$  that reflects the trend as follow:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (6)$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n i Y_i - \sum_{i=1}^n i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n i^2 - (\sum_{i=1}^n i)^2}, \hat{\beta}_0 = \frac{\sum_{i=1}^n i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n i \sum_{i=1}^n i Y_i}{n \sum_{i=1}^n i^2 - (\sum_{i=1}^n i)^2} \quad (7)$$

The obtained trend exists in the whole tracking process. When the features are sufficient, they will continue to rise and fall. Therefore, it is necessary to set the effective number of points  $n = 250$  as the criterion for opening judgment in the tracking process based on experimental experience. Then we use a predictive method to predict the law of changes in environmental conditions. To make the data easier to predict, we use the Savitzky-Golay algorithm [21] to process the data to eliminate excessive fluctuations and use more historical data  $Z_i$  for predicting.  $Z_i (i = 1, \dots, 20)$  are the number of optimized point features in the last twenty frames. In the Savitzky-Golay algorithm, we need to choose the appropriate window size  $m$  and order  $n$  that affect

the degree of data smoothing. According to our experimental experience, we choose  $m = 7$  and  $n = 5$  to deal with fluctuations in  $Z_i (i = 1, \dots, 20)$  can obtain the best smooth data  $Z'_i (i = 1, \dots, 20)$ . In the Savitzky-Golay algorithm that the matrix  $H$  depends only on order  $n$  and window size  $m$  and is independent of the input samples.  $h_{i,m}$  denotes the elements of the matrix  $H$ . The formula for obtaining smoother data is as [21]:

$$Z'_i = \sum_{m=-M}^M h_{i,m} Z[m] \quad (8)$$

Then we perform the second exponential smoothing method prediction on the obtained smoothed data  $Z'_i (i = 1, \dots, 20)$ . To obtain a representative prediction result, we need to choose the appropriate prediction coefficient  $\alpha$  whose size reflects the degree of influence of recent data on the predicted data. After many experiments, we found that setting the prediction coefficient  $\alpha = 0.6$  can obtain a representative prediction result. When  $i = 1$ ,  $Z_1^{(1)} = Z_2^{(2)} = Z'_1$ , and then perform a loop operation to get the predicted value  $\tilde{Z}_{i+1}$  for judgment as follows:

$$Z_i^{(1)} = \alpha Z'_i + (1 - \alpha) Z_{i-1}^{(1)} \quad (9)$$

$$Z_i^{(2)} = \alpha Z_i^{(1)} + (1 - \alpha) Z_{i-1}^{(2)} \quad (10)$$

$$\tilde{Z}_i = \frac{2 - \alpha}{1 - \alpha} Z_i^{(1)} - \frac{\alpha}{1 - \alpha} Z_i^{(2)}, (i = 2, \dots, 20) \quad (11)$$

At the same time, to make the system more stable, we set the stability judgment flag  $\varepsilon_i$  after completing a tracking mode switch. Only when the number of optimized point features of consecutive  $n$  frames is higher than 240 will it be set to be stable, otherwise it is judged to be unstable, and the current tracking state will continue. According to our experiment, when the conditions are met for three consecutive frames, it means that the environment enters a relatively stable state. The final point and line tracking judgment method are as follows.

**Algorithm 1.** Point or Point and Line Method

---

**Input:** Trend flag  $\hat{\beta}_i$ , Predicted value  $\tilde{Z}_{i+1}$ , True value  $Z_i$ , Stability judgment flag  $\varepsilon_i$ .

**Output:** Point Method, Point and Line Method

---

**State: Point Method**

1. When  $Z_i < 250$ , the system enters the judgment of the local trend in the second step, otherwise keeps the Point Method.
2. When  $\hat{\beta}_i > 0$ , it explains that the local trend is rising, and the system continues to keep Point Method; When  $\hat{\beta}_i < 0$ , it explains that the local trend is declining, and the system enters the third step.
3. When  $\tilde{Z}_i > Z_i$ , it proves that the current status is lower than expected, and the system switches to Point and Line Method, otherwise the system keeps the Point Method.

**State: Point and Line Method**

1. When  $Z_i > 240$ , the system enters the judgment of the stability in the second step, otherwise keeps the Point and Line Method.
  2. The system stability flag  $\varepsilon_i$  is used to judge whether the system is in a stable state. When the state is proved to be stable, and the system switches to the Point Method; otherwise the system keeps the Point and Line Method.
- 

## 5 Experimental Validation

In this section, we use the EuRoC datasets [22] and the KITTI datasets [23] to test the effects of our system in indoor and outdoor environments. All experiments run on a laptop computer with i5-9300H 2.40 GHz CPU, 8.00 GB RAM, without GPU.

Our work was modified from stereo ORB\_SLAM [24] and ORB\_Line\_SLAM [18] which obtains better results than stereo PL-SLAM [17] in the dataset test. Our work can enable the system to switch between using point features tracking and point features and line features tracking according to the environment and achieves the same effect as using both point features and line features. To prove this, we compare with the test results of stereo ORB-SLAM and ORB\_Line-SLAM as follow:

### 5.1 EuRoC Dataset

The EuRoC datasets are divided into easy, medium, and difficult according to flying speed, lighting, and environment texture. The datasets provide accurate real poses for the evaluation of the final results.

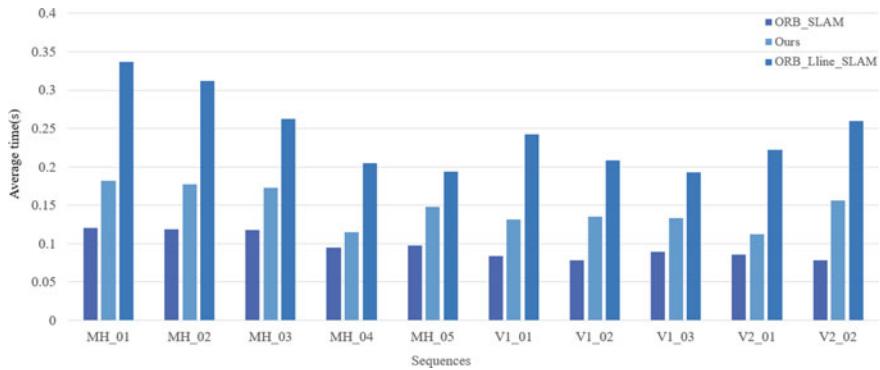
The results are shown in Table 1. We calculate the RMSE of the absolute translation error and label the smallest error as the bold number. Our method performs better than stereo ORB\_SLAM in most sequences. In a series of datasets, our method performs better than ORB\_Line\_SLAM. This is because when the movement is slow,

**Table 1** Comparison of translation RMSE ( $m$ ) on EUROC dataset

Sequence	ORB_SLAM2	ORB_Line_SLAM	Ours
MH_01_easy	0.036639	0.042740	<b>0.035564</b>
MH_02_easy	0.045261	0.047564	<b>0.044164</b>
MH_03_medium	<b>0.034997</b>	0.040741	0.036209
MH_04_difficult	0.052171	0.050491	<b>0.045989</b>
MH_05_difficult	0.055294	0.053610	<b>0.047462</b>
V1_01_easy	0.086873	0.087651	<b>0.086601</b>
V1_02_medium	0.066922	0.063876	<b>0.062852</b>
V1_03_difficult	0.081241	<b>0.065307</b>	0.067131
V2_01_easy	0.065488	0.068203	<b>0.063245</b>
V2_02_medium	0.060031	0.062288	<b>0.055739</b>

and the surrounding environment characteristics are sufficient, the LSD line feature extraction method will affect the tracking accuracy. Therefore, the line features are used when the point feature is insufficient can obtain more accurate pose information. But in V103, ORB\_Line\_SLAM can get the best effect when the camera movement speed is fast, and the spatial characteristics are insufficient.

In Fig. 3, we show a comparison of average tracking time. It can be seen from the data that in an environment with sufficient features and slow camera movement, such as MH01, our method's tracking time is close to ORB\_SLAM. In an environment with lower features, such as MH\_05, V2\_02, our method's tracking time is close to ORB\_Line\_SLAM but still shorter than the tracking time required by ORB\_Line\_SLAM.

**Fig. 3** Comparison of average time(s) on EUROC dataset

**Table 2** Comparison of translation RMSE ( $m$ ) on KITTI dataset

Sequence	ORB_SLAM2	ORB_Line_SLAM	Ours
00	0.915712	0.874138	<b>0.870958</b>
01	5.149767	4.804950	<b>4.309862</b>
02	6.261766	5.514348	<b>4.925226</b>
03	0.291036	0.253860	<b>0.247937</b>
04	0.209263	<b>0.133977</b>	0.169351
05	0.360252	0.400433	<b>0.346013</b>
06	0.533134	0.575652	<b>0.403667</b>
07	0.462235	0.417130	<b>0.412101</b>
08	3.213927	<b>2.929362</b>	3.043945
09	2.804050	1.673850	<b>1.498517</b>
10	1.027561	0.977473	<b>0.745246</b>

## 5.2 KITTI Dataset

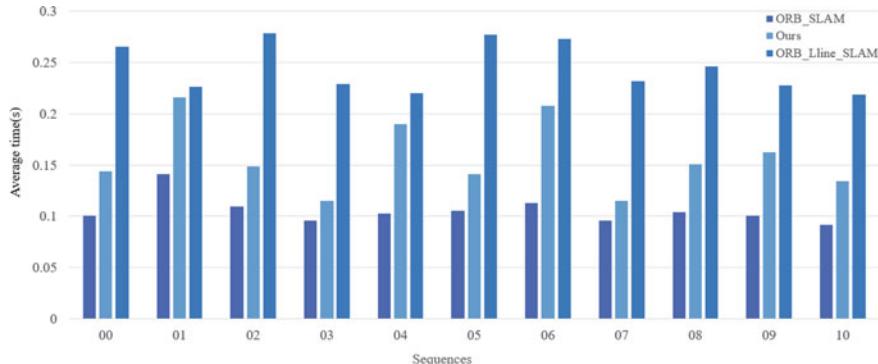
The KITTI datasets are used to test the effect of our system in an outdoor environment. The datasets also provide accurate real poses for the evaluation of the final results.

The results are shown in Table 2. We also calculate the RMSE of the absolute translation error and label the smallest error as the bold number. The outdoor experiment results show that in an outdoor environment, our system can achieve better results than stereo ORB\_SLAM in all the sequences. At the same time, in some urban environments with sufficient texture features, such as 00, it can achieve better results than ORB\_Line\_SLAM. This is because there will be inaccurate line features in an environment with complex textures. But in the urban environment where the surrounding environment features are sparse, and houses have obvious line features, so ORB\_Line\_SLAM performs better in the 04, 08 sequences, but our method can still obtain similar results.

In Fig. 4, We can see that in an urban environment, such as 00, the surrounding environment has sufficient texture features, so the tracking time is close to ORB\_SLAM. On highways environment, such as 01, the surrounding environment has sparse texture features, and the tracking time is close to ORB\_Line\_SLAM, but the required average tracking time is still lower than ORB\_Line\_SLAM.

## 6 Conclusion

Our work is to propose a stereo SLAM system that can reasonably use point features and line features. The system can choose to use different tracking features during the tracking process according to the historical environmental information. When the environment texture features are sufficient, point features are used for tracking; when the environment texture features are not sufficient, point features and line



**Fig. 4** Comparison of average time(s) on KITTI dataset

features are used for tracking. From the experimental results, it can be seen that it is feasible to select features for use based on the historical environment information obtained in the tracking process. Our method was tested on the popular indoor and outdoor public datasets and can obtain higher tracking efficiency and accuracy than the SLAM system that combines both point features and line features.

However, there is still room for improvement in the system. Firstly, the line features extracted during the tracking process are only used in the front-end and tracking process. They are not effectively used in the back-end and loop closure. Secondly, if the environment changes frequently and obviously, the judgment method will occasionally be unstable in tracking. Therefore, we can improve the use of historical environmental information and improve the stability of the judgment method.

## References

1. Cadena, C., Carlone, L., Carrillo, H., et al.: Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. *Trans. Robot.*, pp. 1309–1332 (2016)
2. Milford, M.J., Wyeth, G.F., Prasser, D.: RatSLAM: a hippocampal model for simultaneous localization and mapping. In: *IEEE International Conference on Robotics and Automation*, pp. 403–408. IEEE (2004)
3. Cummins, M., Newman, P.: FAB-MAP: probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.*, pp. 647–665 (2008)
4. Milford, M.J., Wyeth, G.F.: SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights. In: *2012 IEEE International Conference on Robotics and Automation*, pp. 1643–1649. IEEE (2012)
5. Milford, M.J.: Vision-based place recognition: how low can you go? *Int. J. Robot. Res.*, pp. 766–789 (2013)
6. Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: Fast semi-direct monocular visual odometry. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15–22 (2014)
7. Wang, R., Schworer, M., Cremers, D.: Stereo DSO: large-scale direct sparse visual odometry with stereo cameras. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3903–3911 (2017)

8. Scaramuzza, D., Fraundorfer, F.: Visual odometry [tutorial]. *IEEE Robot. Autom. Mag.* **18**(4), 80–92 (2011)
9. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans. Robot.*, pp. 1147–1163 (2015)
10. Montemerlo, M., Thrun, S., Koller, D., et al.: FastSLAM: a factored solution to the simultaneous localization and mapping problem. *Aaaai* (2002)
11. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (IEEE), pp. 225–234 (2007)
12. Rublee, E., Rabaud, V., Konolige, K., et al.: ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision (ICCV), pp. 2564–2571 (2011)
13. Neira, J., Ribeiro, M.I., Tardós, J.D.: Mobile robot localization and map building using monocular vision. In: The 5th Symposium for Intelligent Robotics Systems (1997)
14. Klein, G., Murray, D.: Improving the agility of keyframe-based SLAM. European Conference on Computer Vision (ECCV), pp. 802–815 (2008)
15. Pumarola, A., Vakhitov, A., Agudo, A., et al.: PL-SLAM: Real-time monocular visual SLAM with points and lines. In: 2017 IEEE International Conference on Robotics and Automation, pp. 4503–4508. IEEE (2017)
16. Von Gioi, R.G., Jakubowicz, J., Morel, J.M., et al.: LSD: a fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, pp. 722–732 (2008)
17. Gomez-Ojeda, R., Moreno, F.A., Zuniga-Noël, D., et al.: PL-SLAM: a stereo SLAM system through the combination of points and line segments. *IEEE Trans. Robot.*, pp. 734–746 (2019)
18. Qian, K., Zhao, W., Li, K., et al.: Visual SLAM with BoPLW pairs using egocentric stereo camera for wearable-assisted substation inspection. *IEEE Sens. J.*, pp. 1630–1641 (2019)
19. Strasdat, H., Davison, A.J., Montiel, J.M. M., et al.: Double window optimization for constant time visual SLAM. In: 2011 International Conference on Computer Vision, pp. 2352–2359. IEEE (2011)
20. Gálvez-López, D., Tardos, J.D.: Bags of binary words for fast place recognition in image sequences. *IEEE Trans. Robot.*, pp. 1188–1197 (2012)
21. Schafer, R.W.: What is a Savitzky-Golay filter? [Lecture notes]. *IEEE Signal Process. Mag.*, pp. 111–117 (2011)
22. Burri, M., Nikolic, J., Gohl, P., et al.: The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Res.*, pp. 1157–1163 (2016)
23. Geiger, A., Lenz, P., Stiller C., et al.: Vision meets robotics: the kitti dataset. *Int. J. Robot. Res.*, pp. 1231–1237 (2013)
24. Mur-Artal, R., Tardós, J.D.: Orb-slam2: an open-source slam system for monocular, stereo, and rgbd cameras. *IEEE Trans. Robot.*, pp. 1255–1262 (2017)

# An Intelligent Foreign Substance Inspection Method for Injection Based on Machine Vision



Bowen Zhou , Liang Chen , and Lianghong Wu

**Abstract** The method of intelligent visual detection system for injection realized online high-speed and high-precision detection on foreign substance in the injection. We have researched on the under and back light-given way to obtain sequential images of the injection, put forward adaptive filtering algorithm aimed at small moving targets of the solution to filter out interference of noise points, adopted statistical method of slipping marginal points of window's histogram to position image and detection area, studied on a method that combined two-difference and energy accumulation to extract moving targets, and applied principle of Support Vector Machine to identify foreign substances. A series of experiment demonstrate that the intelligent detection system is able to detect effectively foreign substances in the medical liquid. The detection speed, precision and undetected rate could well meet the needs of a pharmaceutical production line.

**Keywords** Injection · Visual detection · Image processing · Foreign substances recognition

## 1 Introduction

The research on detecting impurities in the injection dated back to a long time ago. The main mechanical operating method is basically high-speed whirl of the injection at first, then scram. At this time, the bottle wall is stationary contrasted to the equipment, but the liquid and the possible impurities in the bottle continue to move due to inertia. Therefore, the impurities can be discriminated against on the basis of their motion characteristics [1]. There are two ways in detection; one is to give the light in the side and receive by optical sensor in the other side according to light scattering and reflection characteristics to judge whether there are impurities that obstruct and influence light irradiation in the liquid, the other is to take sequential images of injection by video camera in accordance with principles of machine vision, and to

---

B. Zhou · L. Chen · L. Wu

School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China

e-mail: [lhwu@hnust.edu.cn](mailto:lhwu@hnust.edu.cn)

extract impurities on the basis of relationship between moving targets and static bottle wall. The first method adopts optical sensor to judge, the bubbles in the bottle are easy to be misjudged as impurities, while the second method identifies impurities on the basis of pictures, which could adopt image processing algorithm to identify that the moving targets is bubble or impurities. Thereby, the research and application is wider [2].

Based on principles of machine vision, image processing algorithm is the key in impurity detection and identification. Some special image processing methods are needed aimed at various types of impurities in the injection. The first is image position, there are image features location method [3] and location algorithm based on marginal point [4]. For the research on feature extraction algorithm, there are target extraction methods based on neural network [5, 6]. Moreover, aimed at target identification, there are particle detection method based on sensor [7], target tracking algorithm based on difference method [8], and target recognition method based on improved PCNN [9]; aimed at more complex background, there are subtraction method based on dynamic background [10], in target tracking method, there are target tracking methods based on frame comparison [11, 12], tracking methods based on mean shift [13], trace tracking method [14], small infrared target tracking method [15] and target tracking method on the basis of three-dimensional features of the image. These methods solved some practical problems and achieved remarkable results.

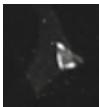
This paper researched on an intelligent visual detection method of injection. By studied an improved adaptive filtering method, put forward statistical position method of slipping marginal points of window's histogram, adopted energy accumulation and twi-difference algorithm to extract moving targets, identified bubble, extracted impurities and separated the unqualified products from production line in the light of target's moving trace, the method will improve greatly the automation level of pharmaceutical safety product line.

In the following paragraphs, Sect. 2 introduces the types of foreign substances in the injection. Section 3 gives the key algorithms of foreign substances detection. Section 4 is devoted to the experiments and analysis of their results. Section 5 gives the conclusion.

## 2 Types of Foreign Substances

In accordance with the information provided by many injection manufacturers, the impurities in the injection include the following types—glass debris, rubber crumb, fiber, dark spot, hair and mosquitoes, among which, glass debris, dark spot and fiber are the most common. Glass debris are mainly results of bottle breakage caused by position and height inaccuracy as filling heads insert bottle during injection filling or caused by excessive extrusion and pressure as the bottles are conveyed during filling or caused by unqualified empty bottle that was crushed and splashed into bottles. Dark spots are mainly the small black residue caused by some medicine carbonation

**Table 1** Foreign substance classification

Type	Glass debris	Fiber	Fluff	black residue	Hair	Mosquito
Color	White	White	White	White	Random	Black
Source	Bottle packing	Bottle collision	Cleaning	Bottle packing	Bottle packing	workers
Size (m n)	50–500	50–200	50–500	50–100	200–1000	500–2000
Sample						

in the bottle because of high-temperature flame when sealed after filling. Fibers are impurities interfused medical solution because of poor filtration before filling.

In accordance with different characteristics of impurities in different illumination condition, we classified the impurities into black impurity and white impurity. Namely, those which can form a clear image when irradiated against the light are black impurities, form a clear image under low-light irradiation are white impurities. The systematical light-given ways will be described in the later section.

Table 1 shows some common impurity types. Aimed at actual injection sample detection, we find that most of the impurities' diameters are between 50 and 200 microns, thereby, they are difficult to be seen by naked eyes. From the analysis on classification features, we could find that there are various types of impurities, such as transparent small glass debris, white fluff, black particles and rubber crumbs of deep color. Therefore, manual detection needs to judge from every angle through different light-given way, which increases the difficulties of detection. Consequently, there is a high omission rate in the existent manual detection by naked eyes.

### 3 Key Algorithms of Foreign Substances Detection

#### 3.1 Image Position and Detection Area Demarcation

Because the image exists plenty of interference and variables in the practical high-speed applications, location algorithm must be characterized with high speed, precision and fault tolerance. The speed of traditional Hough Transform is too slow, the centroid method and optimum fitting method could produce serious errors when the image itself incurs great interference.

It can be found from the above introduction to mechanical structures that the injection bottles just moves right and left but do not whirl in the image. Thus, we just need to find a setpoint. In accordance with the characteristics of injection bottle

images, we choose the side wall and underside whose gray value alteration is remarkable as positioning mark. Based on two lines' intersect of side wall and underside, a setpoint will form. In order to find setpoint quickly, this paper proposed a concise and efficient statistical method of slipping marginal points of window's histogram.

For example, for an injection image  $f(x, y)$ , in order to determine its left and right bottle wall edge, we should utilize formula (Eq. 1) to seek absolute gradient in a horizontal direction.

$$\nabla f(i, j) = |f(i, j) - f(i - 1, j)| + |f(i, j) - f(i + 1, j)| \quad (1)$$

Then we choose appropriate threshold to separate partial marginal spots of injection detected bottles as Fig. 1 shown. Because there are only two obvious edges in vertical direction, we just seek two marginal spots in every line to constitute edge point pairs  $N_{1i}$  ( $i = 1, 2, 3, \dots, n$ ) and  $N_{2i}$  ( $i = 1, 2, 3, \dots, n$ ).  $S_j(X)$  is the sum of marginal spots  $N_{1i}$  and  $N_{2i}$  whose column coordinate is  $j$ .

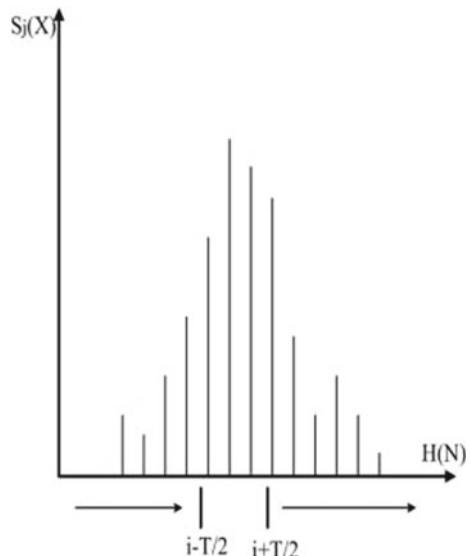
$$S_j(X) = \sum_{N_{11}}^{N_{1n}} H(N) + \sum_{N_{21}}^{N_{2n}} H(N) \quad (2)$$

$S_j(X)$  should be expressed in histogram coordinates as Fig. 2 shown.

**Fig. 1** Location point of the edge



**Fig. 2** Sliding window and histogram of  $X_r$



Supposing the width of sliding window is  $T$ , we can gather statistics  $M_i(S)$ , sum of  $S_j(X)$ , within sliding window.

$$M_i(s) = \sum_{j=i-\frac{T}{2}}^{j=i+\frac{T}{2}} S_j(X) \quad (3)$$

The sliding window glides in the histogram, namely, when value  $i$  is from small to large,  $M(S)$  value changes as well. Therefore,  $M(S)$  acquires two maxima.

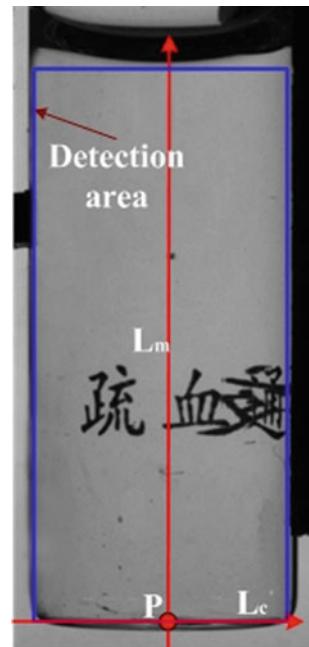
$$X_a = \max M_a(S) \text{ and } X_b = \max M_b(S) \quad (4)$$

The midpoint in sliding window is the setpoint. Because bottle edge is almost a vertical line, bottle wall's positioning line  $L_a$  and  $L_b$  can be acquired. Bottle bottom's positioning line  $L_c$  could be acquired by the same way.

The left and right edges of bottles are almost vertical lines, so the axle wire acquired by distant point is approximately a straight line, axis positioning line  $L_m$  could be acquired. Supposed  $L_m$  and  $L_c$  are intersected at the point  $P(x_a, y_b)$ , it could be bottle's setpoint as Fig. 3 shown.

We complete detection area position in accordance with setpoint and assured detection shape. Histogram sliding window method is able to utilize skillfully statistical value to exclude scattered distribution. But due to great amplitude interference, therefore, we adopt weighted average method to acquire exact values in the sliding window. The experiment demonstrates that there is less demand for marginal point detection for the above algorithm. Even though there are errors in considerable parts

**Fig. 3** Setpoint and detection area



of the marginal detection in the images, the position of target detected in the images could be determined exactly, which is very important in the practical application.

### 3.2 Extraction of Moving Target

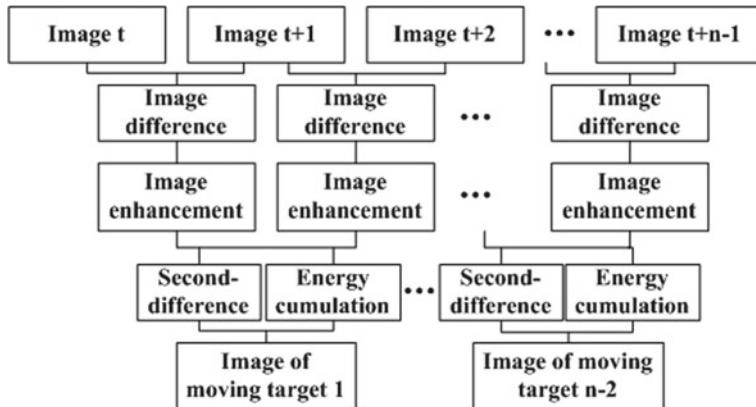
Because the volume of visible foreign substances in the solution is small and their quantities are uncertain, what we need to complete is multi-target tracking to the small impurities in the solution. The system adopts image sequence difference method that combined twi-difference and energy accumulation. The process is shown in Fig. 4.

Suppose acquiring  $n$  sequential images of the solution  $f_t(x, y)$ ,  $f_{t+1}(x, y)$ ,  $f_{t+2}(x, y)$ , ...,  $f_{t+n-1}(x, y)$  ( $n$  is integer more than 10) at time interval  $T$ . Firstly, we calculate by Formula (5) the absolute difference image of the neighbor two frames.

$$d(x, y) = |f_{(t+i+1)}(x, y) - f_{(t+i)}(x, y)| \quad (5)$$

Because the target image is smaller and the energy is lower, we need to utilize the gray difference of the image itself multiplied by enhanced constant  $A$  to make the target have higher energy.

$$P(x, y) = d(x, y) \times A \quad (6)$$



**Fig. 4** The flow chart of moving target obtained

Here, after multiplied A, the gray value of pixel with gray value more than 255 is 255.

The enhanced difference image needs a second difference.

$$D(x, y) = |P_{(t+i, t+i+1)}(x, y) - P_{(t+i-1, t+i)}(x, y)| \quad (7)$$

The energy accumulation of two different images can be calculated by Formula (8) to enhance moving particles against pixel energy.

$$E(x, y) = P_{(t+i, t+i+1)}(x, y) + P_{(t+i-1, t+i)}(x, y) \quad (8)$$

Finally, we subtract  $D(x, y)$  from  $E(x, y)$  so as to obtain gray image of particles.

$$F(x, y) = E(x, y) - D(x, y) \quad (9)$$

Because there are various uncertain types of visible foreign substance in the injection, we need to acquire sequential images of the injection more than ten, and finally obtain image 1, 2, ..., and image n including only moving targets, which could decrease the inaccuracy caused by uncertain factors caused by interference.

### 3.3 Impurities Recognition

The optimal boundary could be obtained through analyzing the similarity and difference between bubble and impurity, as well as, shape, moving speed, track, gray value, and characteristics effect, a great deal of experiments and utilizing the difference in shapes between bubbles and impurities to differentiate.

Upon analysis of shapes, the bubbles are round or similar round, but the shapes of impurities are disorganized. Therefore, the paper classifies and identifies them on the basis of width-to-length ratio of the moving targets. Supposed Area represents moving particle size, L indicates the maximum length of the particles, R is the minimum width of particle, and  $E = L/R$  represents ratio of maximum length and minimum width, if bubbles, the length and width is equal basically, and E is about 1, if impurities, due to their irregularity, E is above 1.5 generally.

As can be seen from the figure that there is a great difference in E value between bubbles and visible foreign substances, we could build SVM model to identify bubbles according to the feature. Judgment whether they are bubbles or not is a problem of binary classification. Supposed that there is statistical data of n moving target,  $x_i$  represents the features of E value. So bubble identification is transformed into a problem to find a suitable classification function which is as follows.

$$f : x_i \rightarrow y_i, y_i \in \{-1, +1\} \quad (10)$$

Firstly, we utilize sample data to train function, and then input experimental lines, using the function to classify. If  $f(x_i) > 0$ , input vector is considered belonging to category of  $y_i = +1$ , that is to say, the moving targets are bubbles. Otherwise, input vector is considered belonging to category of  $y_i = -1$ , namely, the moving targets are visible impurities.

Therefore, training data set  $\Theta$  includes  $x_i$  and  $y_i$  in which the former is input characteristics and the later is classification output results.

$$\Theta = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (11)$$

The classification function adopts the following format.

$$f(x_i) = \text{sgn} \left\{ \sum_{i=1}^n a_i y_i K(x_i, x + b) \right\} \quad (12)$$

Here,  $a_i$  is Lagrange multiplier corresponding to each sample,  $K(x_i, x)$  is inner product function, b is classification threshold,  $x_i$  and  $y_i$  are concentrative training data.

The solving process of  $a_i$  under constraint conditions (13) is as follows.

$$\sum_{i=1}^n y_i a_i = 0 \quad a_i \geq 0, i = 1, 2, \dots, n \quad (13)$$

We acquire the maximum value of the following function for  $a_i$ .

$$Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j K(x_i, x_j) \quad (14)$$

It is easy to get that there is just few of  $a_i$  in the solving is not 0, and the corresponding samples are support vector machines.

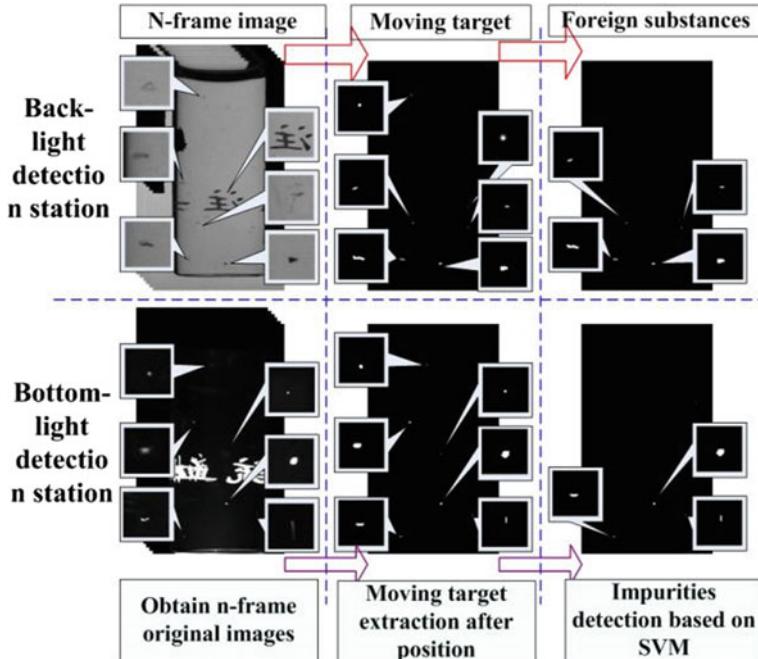
Inner product function  $K(x_i, y_i)$  adopts radial basis function as the follows.

$$K(x_i, x) = \exp\left\{-\frac{|x - x_i|^2}{\sigma^2}\right\}, \text{ (here, } \sigma \text{ is constant)} \quad (15)$$

Categorical threshold  $b$  is acquired by the formula.

$$y_j \left\{ \sum_{i=1}^n a_i y_i K(x_i, x_j) + b \right\} - 1 = 0 \quad (16)$$

Here,  $x_j$  is any support vector, and  $y_j$  is an output result of the support vector. The detection results are as Fig. 5 shown.



**Fig. 5** Result of the impurity detection algorithm

**Table 2** FQ levels and range

FQ levels	0	1	2	4	5	6	7	8	9	10
FQ range	0–0.4	0.5–1.4	1.5–2.4	3.5–4.4	4.5–5.4	5.5–6.4	6.5–7.4	7.5–8.4	8.5–9.4	9.5–10

## 4 Experiments and Analysis

In the world pharmaceutical industries, Knapp-Kushner test program is commonly used to evaluate the detection capability of automatic detection system. The method is acknowledged by European Pharmacopoeia and American FDA based on comparison between detection system capability and election method capacity of any testing in the production or pharmaceutical products. It is considered as the known performance comparison parameter; the algorithm is based on statistical performance evaluation of unqualified products and stands for statistical value of some existent defects. The algorithm evaluation and system testing of this paper are all based on the test program.

In the following experiments, we choose injection of 2 ml produced by some pharmaceutical factory as the experimental objects, adjusting camera and light source to acquire clear images of the injection.

Knapp-Kushner test program is as follows: firstly, we should choose a batch of sample bottles, for example, 250 bottles with number in each one, and then we need to ensure FQ (quality factor) of each bottle by electing x detection worker (5 generally) to detect it ten times repeatedly and totally n times for each batch, and make a record on the detection results. Similarly, we adopt machine to test ten times. Calculation of quality factor is as follows: the detection times of each sample of each person should be added together and is supposed as m, then...

$$FQ = (m/n) \times 10 \quad (17)$$

Each bottle is divided into 11 levels on the basis of its quality as Table 2 shown.

Supposing FQA is quality factor of manual detection, FQB is quality factor of mechanical factor, and the products whose FQ levels is between 7 and 10 is unqualified, then we could obtain...

If  $k$  value is more than 100%, it is indicated that the automatic detection equipment is efficient, that is to say, it is superior to the traditional manual detection. Table 3 represents Knapp-Kushner testing results of an intelligent detection system of injection.

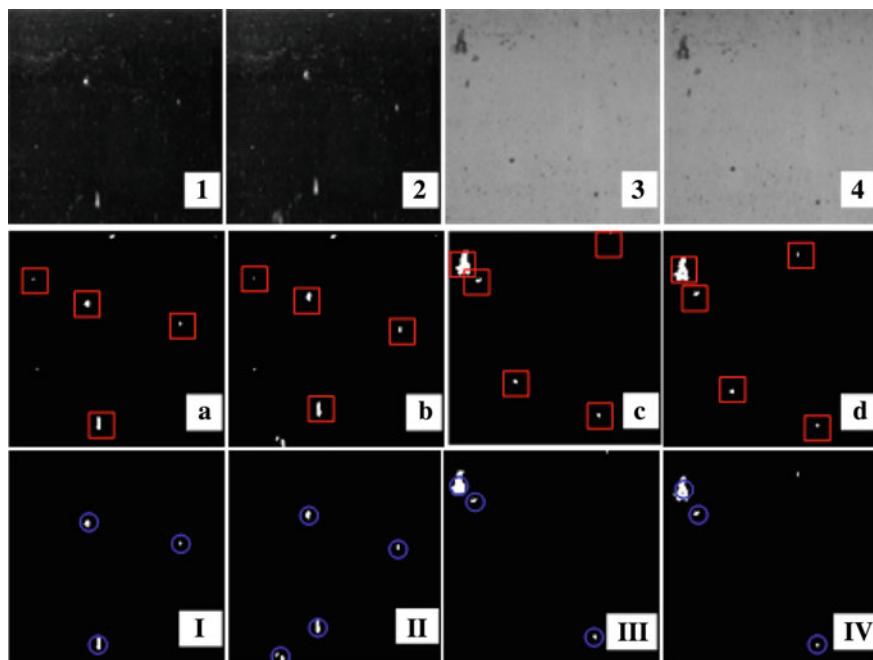
We could test respectively from under light detection station and back light detection station as described in the following.

We adopt amino acid injection samples to carry out algorithm testing after continuous access to images. Because adopting illumination from under light source with black velvet in the back of the bottles, we can obtain the gray images of bright impurities, bubbles and bottle walls with black background. In order to detect performance of the system, we adopt meansift to extract target, and compare it with the method of this paper.

As is shown in Fig. 6 that small pictures 1–4 in the images is primitive sequential images, a–d is results images of meanshift tracking algorithm, I–IV is results images

**Table 3** Knapp-Kushner testing results

Sensitivity	FQA	FQB	$k = FQB/FQA$	Judging criterion	Result
30	437	1142	2.613	$FQB/FQA > 1$	Machine detection is superior to humans



**Fig. 6** Comparative experiment of detection station given light from the bottom

of algorithm of this paper. As is known from the images that meanshift algorithm extract impurities indeed, but it is interfered by bubbles and partial bottle walls. The method in this paper could not only filter the inference of bottle wall, but also bubbles.

## 5 Conclusion

The paper designed and developed an intelligent visual detection method of injection and proved the feasibility of overall structure of the system put forward by the paper. A great deal of online experiments on the machine has verified validity of various detection algorithms put forward by this paper. The detection accuracy meets the detection standards and the detection speed meets the online detection requirements of the automated production line. The system enhances greatly precision of manual light detection, making the detection more accurate and objective improves production efficiency, reduces labor intensity, clean and environmental protection, all of which have great significance in social and economic benefits.

**Acknowledgements** The work is supported by the National youth fund of China (Grant No. 61603132), the Natural Science Foundation of Hunan Province, China (Grant NO. 2020JJ5170) and the Fund of Hunan Provincial Education Department (Grant NO. 18C0299).

## References

1. Fuyu, W., et al.: An opt-electronic method for inspecting foreign particles in injections. *Acta Photonica Sin* **41**(3), 375–378 (2012)
2. Lasheng, Y., Linong, L., Liu, R.: Research on key technology of liquid lamp inspector based on machine-vision. *Comput. Eng. Appl.* **48**(26), 152–154 (2012)
3. Islam, M.J., et al.: Computer vision-based quality inspection system of transparent gelatin capsules in pharmaceutical applications. *Am J Intell Syst* **2**(1), 14–22 (2012)
4. Fei, et al.: A method for positioning mark point on liquid crystal glass based on machine vision. In: CCDC2020, pp. 908–913
5. Beniwal, S., Saini, U., Garg, P., et al.: Improving performance during camera surveillance by integration of edge detection in IoT system. *Int. J. E-Health Med. Commun.* **12**(5), 84–96 (2021)
6. Alvarez-Machancoses, O., Andrés-Galiana, D., Fernández-Martínez, J.L., et al.: Robust prediction of single and multiple point protein mutations stability changes. *Biomolecules* **10**(1), 67 (2019)
7. Kondoh, J., Tada, K.: Continuous measurement of liquid concentration using shear horizontal surface acoustic wave sensors without reference liquid. *Sensors*, pp.1–3 (2017)
8. Juan, L., Yaonan, W., Jie, Z., Bowen, Z.: On-line detection of foreign substances in glass bottles filled with transfusion solution through computer vision. In: 2008 International Conference on Information and Automation (ICIA), pp. 424–429 (2008)
9. Fang, J., Wang, Y., Wu, C.: Binocular automatic particle inspection machine for bottled medical liquid examination. In: Chinese Automation Congress, pp. 397–402. IEEE (2012)

10. Prashant, W.P., Akshay, D., Subrahmanyam, M., et al.: A novel Saliency-based cascaded approach for moving object segmentation. *computer vision and image processing*, pp. 311–322 (2020)
11. Angel, L., Eduardo, J.M., Manuel, O.: Detection and tracking of moving obstacles (DATMO): a review. *Robotica* **38**(5), 761–774 (2020)
12. Yong, L., Shizhong, L., et al.: Kernel stability for model selection in Kernel-based algorithms. *IEEE Trans. Cybern.*, pp. 1–12 (2019)
13. Pallavi, S., Laxmi, K.R., Ramya, N., et al.: Study and analysis of modified mean shift method and Kalman filter for moving object detection and tracking. In: *Proceedings of the Third International Conference on Computational Intelligence and Informatics*, pp. 821–828 (2020)
14. Williams, H.W., Simske, S.J.: Object tracking continuity through track and trace method. *Electron. Imaging*, pp. 2991–2997 (2020)
15. Zhao, L., Tao, H., Chen, W.: Maneuvering target detection based on three-dimensional coherent integration. *IEEE Access* **8**, 188321–188334 (2020)

# A Modified SiamRPN for Visual Tracking



Wei Zhou, Yuxiang Liu, Haixia Xu, and Zhihai Hu

**Abstract** Siamese network based trackers have achieved state-of-the-art performance on multiple benchmarks. SiamRPN can predict the size of target thanks to RPN module. This paper proposes a modified SiamRPN based on IoU, under the framework of SiamRPN, Siamese feature extraction, and proposal generation for target, followed by the loss minimization. Aiming to the loss of both the classification branch and regression branch, we introduce IoU between GT box and the anchor into the regression loss function to form the joint optimization of IoU&smooth L<sub>1</sub> norm, which is useful to refine the tracking target box prediction. Then, we define IoU between GT box and the predicted box to weight positive samples. Weighted positive samples establish the connection between the classification branch and regression branch, which is helpful to eliminate the inconsistency in the optimal prediction of two branches. Experimental evaluations on the datasets OTB2013, OTB2015, demonstrate that compared with the state-of-the-art tracker such as SiamFC, SiamRPN and other algorithms, our proposed tracker achieves higher tracking accuracy and stronger robustness in most challenges of the tracking situation.

**Keywords** Target tracking · Siamese network · Intersection over union (IoU)

## 1 Introduction

In recent years, deep learning [1] has been leading the progress of visual tasks, such as target tracking [2], image segmentation [3] and target detection [4] and others. We focus on the target tracking, and also pay close attention to other tasks for they promote each other.

Trackers based on Siamese network have attracted many researchers thanks to their balance of speed and accuracy. Tao et al. [5] propose Siamese instance search for tracking (SINT), which adopts Siamese network structure, matching candidate image patches with multi-scale and the target patch. Then, Bertinetto et al. [6] design

---

W. Zhou · Y. Liu (✉) · H. Xu · Z. Hu  
Xiangtan University, Xiangtan 411105, China

the full convolution Siamese network, named SiamFC, which measures the similarity between the search image features and the target image feature through correlation convolution, and formulate the target tracking into the problem of the image matching. SiamRPN [7] introduces region proposal network (RPN) into the Siamese network, and utilizes the anchor mechanism of the object detect task [8] to predict the size of target. Therefore, a boundary box regression branch and a classification branch are added to SiamFC to discriminate the target and bound the target candidate region. Dsiam [9] explores a dynamic Siamese network to learn object appearance changes and background suppression online, and trains them with continuous video frames. DasiamRPN [10] uses the detection dataset to expand the positive samples and the difficult negative samples, and designs the interference perception module to distinguish the real target from the disturbance, which improves the generalization of the tracker.

SiamRPN implements the size prediction of the target by introducing RPN module, but several aspects are to be modified.

Firstly, the regression branch in SiamRPN is optimized by  $L_1$  norm loss, so the prediction of bounding box is not accurate [11, 12].

Secondly, SiamRPN filters positive and negative samples through the Intersection over Union (IoU) ratio between anchor and the Ground Truth (GT) bounding box, which leads to low discrimination among positive samples.

Finally, the classification branch is separate from the regression branch in the introduced RPN module, which may not lock the same candidate target patch in the optimal prediction of the two branches.

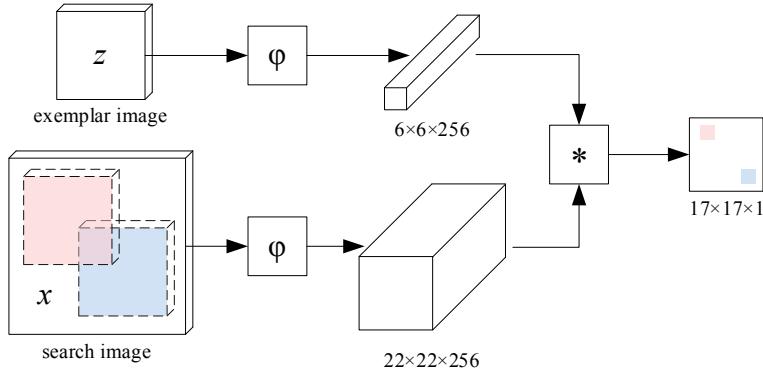
In this paper, we propose a modified SiamRPN based on IoU. Under the framework of SiamRPN, we introduce IoU between GT box and anchors into the loss function to refine the regression prediction box, and define IoU between GT box and predicted box to weight positive sample for distinguishing each other, and positive samples based on IoU establish the connection between the classification branch and regression branch. Tracking experiments are carried out on OTB2013 [12], OTB2015 [13] test datasets to verify the feasibility and effectiveness of the proposed tracker.

The remainder of this paper is organized as follows. Section 2 discusses the principle of Siamese network. A modified SiamRPN for visual tracking is proposed in Sect. 3. In Sect. 4 experiments and discussion are given. The final section presents conclusion as well as future work.

## 2 Siamese Network

The classic Siamese network used in the tracking task is shown in Fig. 1. It formulates the problem of target tracking into the matching one between images.

Siamese networks apply an identical transformation  $\varphi$  to both exemplar image  $z$  and candidate image  $x$ , and measure the similarity between their representations by cross-correlation Layer as follows.



**Fig. 1** Siamese network

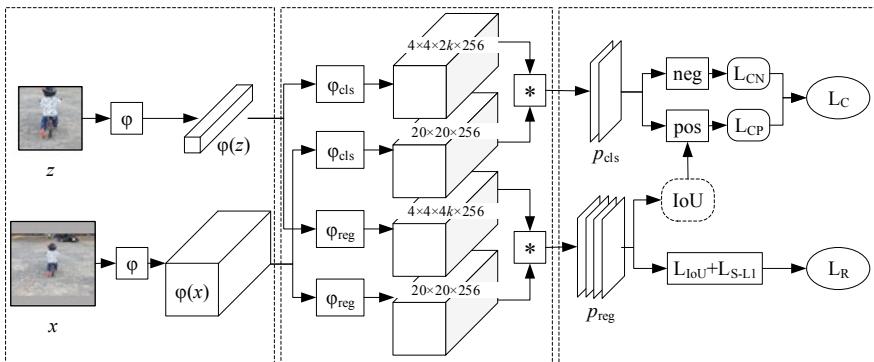
$$f(z, x) = \varphi(z) * \varphi(x) + b \quad (1)$$

where  $b$  is a bias at every location of score map.

The similarity measure function  $f(z, x)$  is learned to evaluate the similarity between the exemplar features and the candidate features, so as to obtain the similarity response score map that shows a high score if the two images depict the same object and a low score otherwise.

### 3 The Proposed Method

In this section, we propose the modified Siamese-RPN based on IoU, illustrated in Fig. 2. Under the framework of SiamRPN, we introduce IoU between GT box and



**Fig. 2** A modified Siamese-RPN framework

anchors into the loss function to refine the regression prediction box, and define IoU between GT box and predicted box to weight positive sample for distinguishing each other, and positive samples based on IoU establish the connection between the classification branch and regression branch.

We divide it into four parts a Siamese feature extraction module, region proposal module, bounding box regression, and foreground–background classification.

### 3.1 Siamese Feature Extraction Module

Siamese feature extraction module is to map images into feature representation domain. As is shown on the left block of Fig. 2, it consists of two branches, one is for the feature extraction of exemplar image, which is from the historical frame, we denote input  $z$ , output  $\varphi(z)$ . The other is for search image which is from the current frame, we denote input  $x$ , output  $\varphi(x)$ .

They share the learnable network  $\varphi$ , which adopts a full convolution network layer of Alexnet [14]. That is, The input  $z$  with  $127 \times 127$ , got by center cropping and input  $x$  with  $255 \times 255$  got, in the same manner, are fed into Siamese module for feature extraction.

### 3.2 Region Proposal Module

The region proposal module is to obtain proposal generation for tracking targets. As is shown in the middle block of Fig. 2, it has two Siamese convolution networks  $\varphi_{cls}$ ,  $\varphi_{reg}$ , used for foreground & background classification branch and target bounding box regression branch, respectively, and each of which is matched with a supervision section.

In order to get the feature in the identical representation domain, the two Siamese convolution networks  $\varphi_{cls}$ ,  $\varphi_{reg}$ , are applied to features  $\varphi(z)$ ,  $\varphi(x)$ , respectively, and output  $\varphi_{cls}[\varphi(z)]$ ,  $\varphi_{cls}[\varphi(x)]$  for classification, and  $\varphi_{reg}[\varphi(z)]$ ,  $\varphi_{reg}[\varphi(x)]$  for regression.

Then we perform the cross-correlation on the classification branch and the regression branch as below

$$\begin{aligned} p_{cls} &= \varphi_{cls}[\varphi(z)] * \varphi_{cls}[\varphi(x)] \\ p_{reg} &= \varphi_{reg}[\varphi(z)] * \varphi_{reg}[\varphi(x)] \end{aligned} \quad (2)$$

Here,  $\varphi_{cls}[\varphi(z)]$  and  $\varphi_{reg}[\varphi(z)]$  server as convolution kernel,  $\varphi_{cls}[\varphi(x)]$  and  $\varphi_{reg}[\varphi(x)]$  server as input signal in the cross-correlation layer.

Anchor mechanism is introduced to the tracking task. If there are  $k$  anchors, classification prediction  $p_{\text{cls}}$  2  $k$  channels, and regression prediction  $p_{\text{reg}}$  output 4  $k$  channels.

### 3.3 Loss Function

In this section, as is shown on the right block of Fig. 2, we introduce IoU into loss function, and re-formulate the loss function of the regression branch and classification branch, respectively, as the following subsections.

We apply the strategy from SiamRPN [7] to pick positive and negative training samples: In terms of IoU between anchors and Ground truth box of target, positive samples are defined as anchors which has  $\text{IoU} > 0.6$ . and negative samples are defined as anchors which have  $\text{IoU} < 0.3$ . We limit at most 16 positive samples and totally 64 samples from one training pair, and optimize the loss function of bounding box regress on the positive samples, and loss function of classification on total samples. We set 5 anchors with the same area and the aspect ratios [0.33, 0.5, 1, 2, 3].

**Regression Loss.** It is not so effective to use only  $L_1$  norm loss for the optimizer of the bounding box regression in the SiamRPN.

According to the works [11, 15] survey, IoU loss is one of the most effective evaluation, and is more accurate than that of the  $L_n$  norm loss in the bounding box regression. However, IoU loss has the difficulties of the highly nonlinear, multi-degree of freedom and the multiple zero gradient regions [16], it is hard to optimize IoU loss. Meanwhile, parameter imbalance exists in RPN module [17]. It is further hard to optimize IoU loss of the RPN network. I guess it may be the main reason why SiamRPN doesn't directly use IoU loss.

Here, we develop the bounding box regression prediction loss based on the joint optimization of IoU loss and smooth  $L_1$  norm loss.

In order to overcome the difficulty of IoU loss, we optimize only the IoU loss of the best positive sample, and optimize smooth  $L_1$  loss on the other positive samples. It is noted that the best positive sample is defined as the anchor that has the max IoU.

At the same time, the best positive sample is located in the central region. IoU loss will play a less important role in training process if only being optimized on the best positive sample. We illustrate the joint optimization of IoU loss & smooth  $L_1$  norm loss processing in Fig. 2.

To begin with input, exemplar image  $z$  is got by center cropping. The search image  $x$  is got by cropping at a new center, which is shifted with random pixels. Then inputs  $z, x$  are fed into Siamese module and RPN module to output the prediction. The loss of target bounding box regression based on IoU & smooth  $L_1$  is given as

$$L_R = L_{\text{best}} + \sum_{i \in \text{pos}} L_{S-L_1}(p_{\text{reg}}^{(i)}) \quad (3)$$

where pos is all of positive samples except the best positive sample.  $L_{S-L_1}$  is smooth  $L_1$  loss, which is computed as SiamRPN [7]. The loss defined as on the best positive sample  $L_{\text{best}}$  is formulated by

$$L_{\text{best}} = 1 - I_{\text{IoU}}(b_{\text{reg}}^{(\text{best})}, \text{gt}_{\text{reg}}) + R_{\text{penalty}}(b_{\text{reg}}^{(\text{best})}, \text{gt}_{\text{reg}}) \quad (4)$$

where  $\text{gt}_{\text{reg}} = \{(x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}})\}$  is GT target bounding box,  $b_{\text{reg}}^{(\text{best})} = \{(x_b, y_b, w_b, h_b)\}$  is the predicted target bounding box on the best positive sample.  $I_{\text{IoU}}(b_{\text{reg}}^{(\text{best})}, \text{gt}_{\text{reg}})$  is the IoU between  $\text{gt}_{\text{reg}}$  and  $b_{\text{reg}}^{(\text{best})}$ . Penalty term of IoU loss  $R_{\text{penalty}}$  describes a constraint on bounding box, and it is calculated as Ref. [18]

$$R_{\text{penalty}}(b_{\text{reg}}, \text{gt}_{\text{reg}}) = \frac{\rho^2(b_{\text{reg}}, \text{gt}_{\text{reg}})}{C^2} + \alpha v \quad (5)$$

where  $\rho(b_{\text{reg}}, \text{gt}_{\text{reg}})$  is Euclidean distance between  $\text{gt}_{\text{reg}}$  and  $b_{\text{reg}}^{(\text{best})}$ . The weight coefficient  $\alpha = \frac{v}{(1 - I_{\text{IoU}}(b_{\text{reg}}, \text{gt}_{\text{reg}})) + v}$ .  $v$  is used to measure the similarity of length-width ratio between the ground truth box and the predicted box, and computed by

$$v = \frac{4}{\pi^2} \left( \arctan \frac{w_{\text{gt}}}{h_{\text{gt}}} - \arctan \frac{w_b}{h_b} \right)^2 \quad (6)$$

From Eqs. (4) to (6), it can be seen that IoU loss keeps the target accuracy to the most extent in such aspects of intersection, length-width and center distance.

**Classification Loss.** SiamRPN picks positive and negative samples based on IoU between GT box and anchors. There is only one target in each image for the single target tracking task, so the positive samples are all from the same target. It is hard to determine which positive sample approaches more to the true target when their IoU is close to each other.

On the other hand, regression branch is separate from classification branch in SiamRPN, which may not lock the same candidate target patch in the optimal prediction of the two branches.

In this paper, we define weight coefficients for positive samples based on IoU between GT bounding box  $\text{gt}_{\text{reg}}$  and the predicted bounding boxes  $b_{\text{reg}}^{(\text{pos})}$ , which are returned by regression branch.

The weight is used to distinguish sampled positive samples from each other. Consequently, these weighted positive samples bridge classification prediction and regression prediction. It is helpful to overcome the inconsistency by establishing the connection between classification prediction and regression prediction.

Then we formulate the classification loss on negative samples and weighted positive samples as below

$$L_C = L_{\text{CP}} + L_{\text{CN}} \quad (7)$$

The classification loss on weighted positive samples is given by

$$L_{CP} = \sum_{i \in pos} L_{CE}\left(\eta_{scale} \cdot I_{IoU}\left(b_{reg}^{(i)}, gt_{reg}\right) \cdot p_{cls}^{(i)}, gt_{cls}^{(i)}\right) \quad (8)$$

where  $L_{CE}(x, y)$  is cross-entropy loss function,  $gt_{cls}^{(i)}$   $p_{cls}^{(i)}$  are the ground truth and predicted classification logits of the  $i$ th positive sample, respectively.  $I_{IoU}\left(b_{reg}^{(i)}, gt_{reg}\right)$  is weight coefficient for the  $i$ th positive sample.

All of positive samples weights is scaled by a scalar  $\eta_{scale}$  to reduce the stochastic volatility of regression prediction. Based on IoU and prediction,  $\eta_{scale}$  is defined as

$$\eta_{scale} = \frac{\sum_{i \in pos} p_{cls}^{(i)}}{\sum_{i \in pos} I_{IoU}\left(b_{reg}^{(i)}, gt_{reg}\right) p_{cls}^{(i)}} \quad (9)$$

The classification loss on negative samples is given as

$$L_{CN} = \sum_{i \in neg} L_{CE}\left(p_{cls}^{(i)}, gt_{cls}^{(i)}\right) \quad (10)$$

where neg denotes negative samples.

Finally, the total loss function on two branches is given as

$$L_{SUM} = L_R + L_C \quad (11)$$

## 4 Experiments

In this section, we evaluate our proposed algorithm by conduct experiments on benchmark datasets OTB2013 [12], OTB2015 [13]. All the tracking results ensure a fair comparison.

### 4.1 Parameter Settings and Implementation Details

**Parameter settings.** All of experiments run on Ubuntu 18.04, Python3.6.12 and Pytorch1.6.0 platform with an Intel Xeon Gold 5122 CPU and a GeForce RTX 2080Ti GPU, memory 16 GB.

These parameters of Siamese module and RPN module are obtained by optimizing loss function in Eq. (11) with Stochastic Gradient Descent (SGD). We perform 50 epochs with mini-batch 32, the learning rate decreased  $10^{-2}$  to  $10^{-6}$  at each epoch.

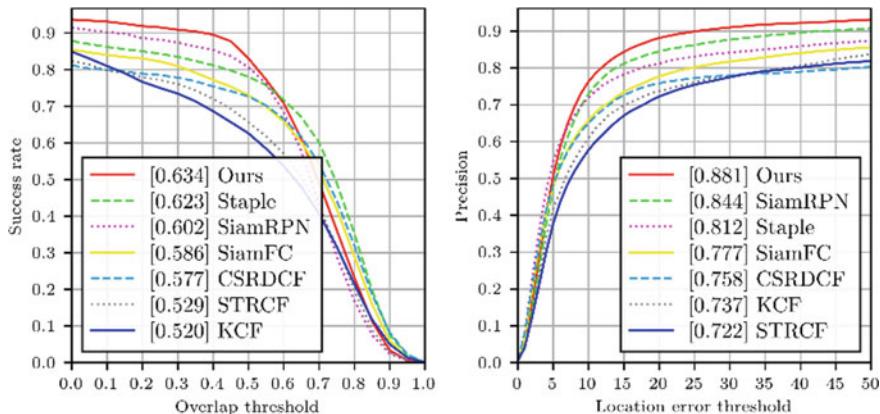
**Implementation details.** During offline training phase, we train our proposed Siamese-IoU through end-to-end on datasets GOT10K [19] and on YouTube-Bounding-Boxes [20]. During online tracking phase, there is no online adaptation since we formulate online tracker as one-shot detector.

## 4.2 Quantitative Analysis

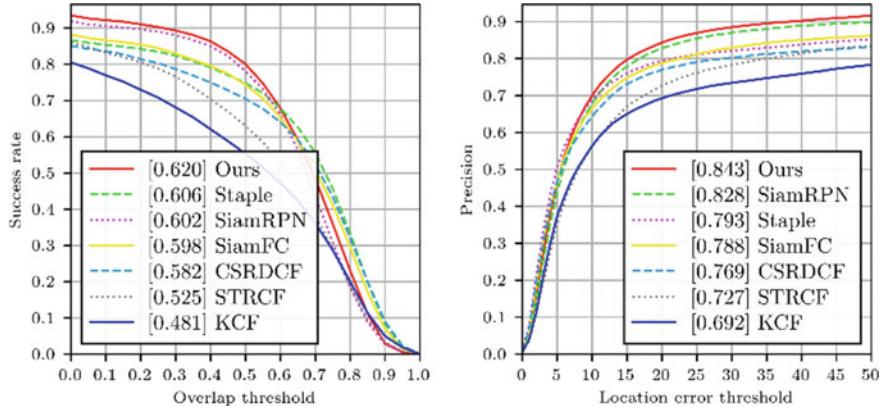
Our proposed tracker is evaluated and compared with top other trackers SiamFC [6], SiamRPN [7], Staple [21], KCF [22], CSRDCF [23], STRCF [24]. Here, trackers SiamFC and SiamRPN are trained offline with the above parameter settings and implementation details, and tracked online with their default hyperparameters.

**Evaluation criteria.** (1) precision, report the ratio of successful frames which Euclidean distance between the center of the predicted bounding box and the center of the ground truth is less than the given threshold  $\tau$  ( $\tau$  is set to 20 pixels) to the total number of video frames. (2) success rate: report the ratio of the number of frames whose overlap score is greater than the given threshold ( $\tau$  is set to 0.5) to the total number of video frames.

**Result on OTB2013.** OTB2013 datasets contain 50 video clips. The performance is evaluated in terms of success plot and precision plot. The tracking results are reported on the test sets of OTB2013 in Fig. 3. It can be seen that the tracker Ours achieves an average precision 88.1% and a success rate of 63.4%. Tracker Ours is superior to other trackers SiamRPN, SiamFC, Staple, KCF, CSRDCF, STRCF.



**Fig. 3** Success plots and precision plots on OTB2013



**Fig. 4** Success plots and precision plots on OTB2015

Compared with top tracker SiamRPN, tracker ours increases by 3.2% and 3.7% in precision and success, respectively.

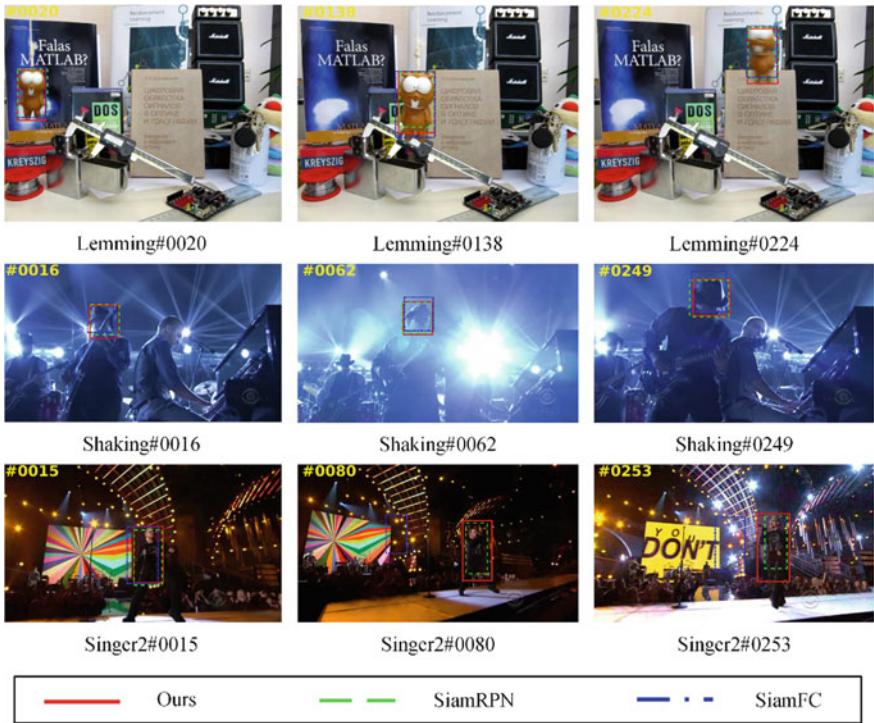
**Result on OTB2015.** OTB2015 datasets contain 100 video clips. The tracking results of the success plot and precision plot are illustrated in Fig. 4 on the test sets of OTB2015. It can be seen that the tracker ours achieves average precision 84.3% and success rate 62.0%. Tracker Ours is superior to other trackers SiamRPN, SiamFC, Staple, KCF, CSRDCF STRCF. Compared with top tracker SiamRPN, tracker Ours increase by 1.5% and 1.8% in precision and success, respectively.

To sum up, our proposed tracker (SiamIoU) outperforms significantly overSiamRPN, SiamFC and others in accuracy and EAO.

### 4.3 Qualitative Analysis

To intuitively evaluate and demonstrate Tracker Ours, we visualize the tracking comparison with SiamRPN, SiamFC on the following challenging clips from OTB2013, Lemming, Shaking, Singer2 and Ironman in Fig. 5. We give a brief qualitative analysis of the tracking visualization.

For the challenges of the Background Cluster (BC), Illumination Variation (IV), as can be seen in the sequence of Shaking, Singer and Lemming. The tracker Ours shows better robustness to IV than SiamRPN and SiamFC, for instance, the results of the frames that happen to flashlight on the clip Shaking. Our tracker bounds the target well thanks to the introduction of the IoU refine.



**Fig. 5** Comparison of the tracking results of Ours with SiamRPN and SiamFC

## 5 Conclusion

In this paper, we propose a modified Siamese region proposal network based on the IoU, It is end-to-end offline trained on datasets GOT10K and YouTube Bounding-Boxes by applying box refinement procedure. In the inference phase, Our tracker is formulated as a local one-shot detector, and outperform SiamRPN and other trackers on datasets OTB2013, OTB2015.

**Acknowledgements** This work was supported by the Science and Technology Plan Project of Hunan Province (2016TP1020), open fund project of Hunan Provincial Key Laboratory of Intelligent Information Processing and Application for Hengyang normal university (IIPA20K04).

## References

1. Zhang, R., Li, W., Mo, T., et al.: Review of deep learning. *Inf. Control* **47**(4), 385–397 (2018)
2. Hou, Z., Dai, B., Hu, D., et al.: Robust visual tracking via perceptive deep neural network. *J. Electron. Inf. Technol.* **38**(7), 1616–1623 (2016)

3. Li, D., Zhang, Z.: Improved U-Net segmentation algorithm for the retinal blood vessel images. *Acta Opt. Sin.* **40**(10), 101–110 (2020)
4. Guo, Z., Song, P., Zhang, Y., et al.: Aircraft detection method based on deep convolutional neural network for remote sensing images. *J. Electron. Inf. Technol.* **40**(11), 2684–2690 (2018)
5. Tao, R., Gavves, E., Smeulders, A.: Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 1420–1429 (2016)
6. Bertinetto, L., Valmadre, J., Henriques, J., et al.: Fully-convolutional Siamese networks for object tracking. In: European Conference on Computer Vision, Amsterdam, Netherlands, pp. 850–865 (2016)
7. Li, B., Yan, J., Wu, W., et al.: High performance visual tracking with Siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 8971–8980 (2018)
8. Zhu, Z., Wang, Q., Li, B., et al.: Distractor-aware Siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 101–117 (2018)
9. Ren, S., He, K., Girshick, R., et al.: Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2016)
10. Guo, Q., Feng, W., Zhou, C., et al.: Learning dynamic Siamese network for visual object tracking. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 1763–1771 (2017)
11. Rezatofighi, H., Tsoi, N., Gwak, J.Y., et al.: General-ized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, pp. 658–666 (2019)
12. Wu, Y., Lim, J., Yang, M.: Online object tracking: a benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, pp. 2411–2418 (2013)
13. Wu, Y., Lim, J., Yang, M.: Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1834–1848 (2015)
14. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017)
15. Yu, J., Jiang, Y., Wang, Z., et al.: Unitbox: an advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, Netherlands, pp. 516–520 (2016)
16. Tychsen-Smith, L., Petersson, L.: Improving obj-ect localization with fitness nms and bounded iou loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 6877–6885 (2018)
17. Li, B., Wu, W., Wang, Q., et al.: Siamrpn++: evolution of Siamese visual tracking with very deep net-works. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019, pp. 4282–4291 (2019)
18. Zheng, Z., Wang, P., Liu, W., et al.: Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, pp. 12993–13000 (2020)
19. Huang, L., Zhao, X., Huang, K.: Got-10k: a large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* (2019)
20. Real, E., Shlens, J., Mazzocchi, S., et al.: Youtube-boundingboxes: a large high-precision human-annotated data set for object detection in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 5296–5305 (2017)
21. Bertinetto, L., Valmadre, J., Golodetz, S., et al.: Staple: complementary learners for real-time tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, pp. 1401–1409 (2016)
22. Henriques, J., Caseiro, R., Martins, P., et al.: High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 583–596 (2014)

23. Lukezic , A., Vojir, T., ĆehovinZajc, L., et al.: Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017, pp. 6309–6318 (2017)
24. Li, F., Tian, C., Zuo, W., et al: Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 4904-4913 (2018)

# Unsupervised Person Re-identification via Multi-branch Network



Xiaobin Wang · Baodi Liu · and Weifeng Liu

**Abstract** Currently, the mainstream unsupervised method uses a clustering algorithm to cluster unlabeled data and generate labels for training samples based on the clustering results. The critical step is to obtain the features. Most unsupervised methods are based on single-branch networks. Still, multi-branch networks that fusion local and global features have been proved to improve supervised effectively. This paper studies the problem of unsupervised person re-identification using the multi-branch network. The goal is to extract more reliable feature representation through the multi-branch network and obtain a model with a more vital distinguishing ability. Specifically, an OSNet-based multi-branch network simultaneously extracts global, local, and channel features. It applies a bottom-up clustering algorithm based on the hierarchy to create labels for unlabeled training samples and uses Softmax-Triplet joint loss to optimize the model. We verify the proposed method on two benchmark re-ID datasets, such as the Market-1501 and DukeMTMC-reID datasets. Compared with the baseline method, the accuracy of Rank-1 and mAP are improved by 5.5 and 12.6% on Market-1501, and 5.5 and 7.6% on DukeMTMC-reID.

**Keywords** Multi-branch network · Cluster · Joint loss

## 1 Introduction

Person re-identification (re-ID) [1] is a retrieval problem, and the target is to retrieve the target across cameras in the gallery based on the given query target. Traditional person re-identification methods include KISSME [2] and DNS [3]. Benefit from the robust performance of the convolutional neural network (CNN), the supervised person re-identification methods [4–9] have achieved excellent performance on the Market-1501. Still, supervised methods require manual annotation of the datasets, which is time-consuming and laborious. In practical applications, the amount of data is tens of thousands, which cannot be labeled, so the supervised method is unsuitable

---

X. Wang · B. Liu · W. Liu

College of Control Science and Engineering, China University of Petroleum (East China),  
Qingdao, China

e-mail: [liubaodi@upc.edu.cn](mailto:liubaodi@upc.edu.cn)

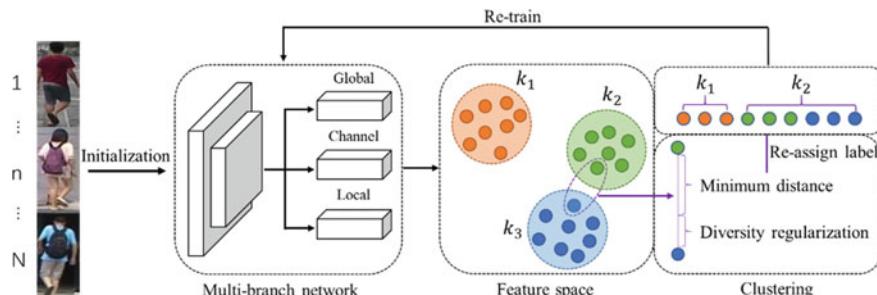
for practical applications. The unsupervised person re-identification method [10–13] does not require annotating the dataset, and it is easier to apply to practical applications, so it has received widespread attention.

The difference between unsupervised and supervised is whether the training data have labels. To solve this problem, several works [14–16] regard it as a transfer task, pre-training on labeled training sets, and then perform domain alignment or label migration. These methods have achieved particular success, but they always use labeled datasets not to be regarded as entirely unsupervised learning. Under the condition of completely unsupervised learning, researchers propose to use the similarity among the same samples and the dissimilarity among different samples as the supervised information. Then, they generate pseudo labels for unlabeled data through clustering algorithms based on hierarchical [13] or density [17]. But these methods only focus on the global or specific regional feature representation.

This paper adopts a multi-branch network architecture that can simultaneously extract global, local, and channel features and comprehensively considers three scales' feature representation. Figure 1 shows the subject framework of the algorithm. The initial stage is to treat each training set sample as a separate class and assign an individual pseudo label. The generated pseudo labels are applied to initialize the network. In the second stage, the hierarchical clustering algorithm is applied to the multi-scale features to generate pseudo labels for the unlabeled samples. In the third stage, use the generated pseudo labels as supervised information, and apply the Softmax-Triplet joint loss function to optimize the network. The latter two stages are alternated until the network performance reaches the highest on the test set.

The main contributions of this paper are as follows,

- We propose a multi-branch network to extract multi-scale features simultaneously to obtain a more reliable feature representation.
- We propose to utilize the Softmax-Triplet joint loss to optimize the multi-branch network.
- We verify the proposed multi-branch approach on two benchmark datasets, and the experimental results prove the effectiveness of this method.



**Fig. 1** The framework of the proposed algorithm

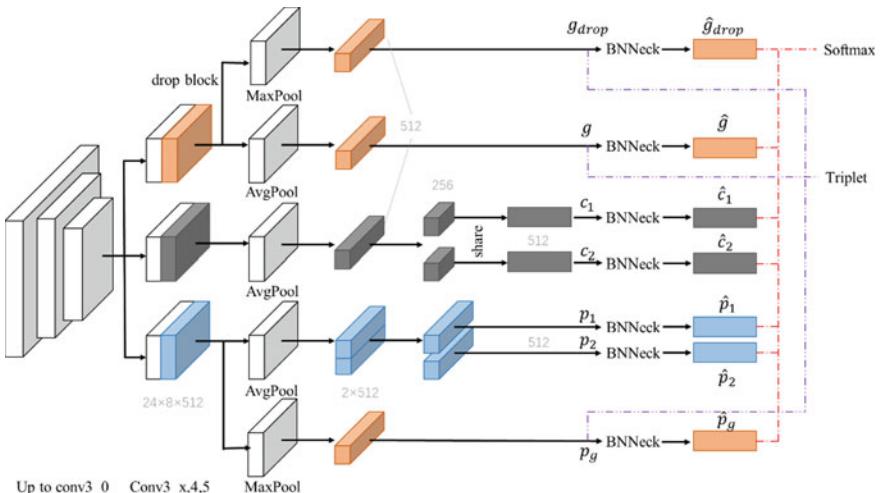
## 2 Methodology

### 2.1 Preliminary

Given an unlabeled training set  $X^t = \{x_1^t, x_2^t, \dots, x_{N_t}^t\}$  with  $N_t$  samples, a labeled query set  $X^q = \{x_1^q, x_2^q, \dots, x_{N_q}^q\}$  with  $N_q$  samples and a labeled gallery set  $X^g = \{x_1^g, x_2^g, \dots, x_{N_g}^g\}$  with  $N_g$  samples. The target is to obtain an effective embedding function  $\varphi = (\theta; x_i)$  through the unlabeled training set  $X^t$ . This embedding function is used to obtain the feature embeddings of the query set  $X^q$  and gallery set  $X^g$  samples, and to minimize the Euclidean distance between any query sample  $x_i^q$  and the similar sample  $x_j^g$  in the gallery to obtain the best retrieval results.

### 2.2 Network Architecture

At present, the multi-branch architecture has achieved excellent results in supervised methods and proved superior performance compared with the original single-branch architecture. In this paper, LightMBN [18], a multi-branch architecture similar to MGN [19] and SCR [20], was used as the network architecture to extract multi-scale feature representations. This architecture uses OSNet [21] designed for re-ID tasks as the backbone network and creates a three-branch architecture based on global features, partial features, and channel features. As shown in Fig. 2, the network architecture comprises three branches: global, channel, and local.



**Fig. 2** Multi-branch network architecture

The input of the model is the sample  $x^t$  in the unlabeled training set  $X^t$ , and the size is adjusted to  $384 \times 128$ . The model is divided into a shared layer and an individual layer. The shared layer is composed of the layers before OSNet conv3\_0, shared by three branch networks. The input  $x^t$  enters the individual layer after passing through the shared layer. The individual layer comprises three-branch networks, and each branch is composed of the layer after OSNet conv3\_0, the pooling layer, and the BNNeck layer. They are independent of each other, and their weights are not shared.

In the global feature branch, two global feature representations are obtained. The input  $x^t$  passes through the shared layer and the unique conv3\_x, conv4, and conv5 layers of the global branch to obtain a  $24 \times 8 \times 512$ -dimensional tensor. The initial  $24 \times 8 \times 512$ -dimensional tensor is used as the input of a dropout block. The dropout block will remove the regions with the highest activation level in the tensor and reserve the areas with low activation levels to force the network to pour attention into the areas with less discrimination. After the dropout block, 2D max-pooling is applied to the tensor to obtain the first 512-dimensional global feature representation  $g_{drop}$ . The second 512-dimensional global feature representation  $g$  is obtained by applying 2D average pooling to the initial  $24 \times 8 \times 512$ -dimensional tensor.

In the channel feature branch, the input  $x^t$  passes through the shared layer and the unique conv3\_3, conv4, and conv5 layers of the channel branch to obtain a  $24 \times 8 \times 512$ -dimensional tensor. Average pooling is applied to the initial tensor to obtain a 512-dimensional vector. After splitting it into 256-dimensional vectors, use the share layer to expand into two 512-dimensional channel features representation  $c_1$  and  $c_2$ . The share layer comprises  $1 \times 1$  convolution, batch normalization, and ReLU activation function, and two 256-dimensional channel branches share the weight.

In the local feature branch, we use average pooling on the  $24 \times 8 \times 512$ -dimensional tensor obtained through the shared layer and the individual layer to obtain a  $2 \times 512$ -dimensional vector, then spilled it to obtain two 512-dimensional local features  $p_1$  and  $p_2$  Which are representing the upper and lower parts of the sample. In addition, the third 512-dimensional global feature representation  $p_g$  is obtained by using max pooling on the initial  $24 \times 8 \times 512$ -dimensional tensor.

After obtaining the 512-dimensional representation of each branch, use the BNNeck [22] block to optimize the feature representation further. The BNNeck block is composed of batch normalization and a fully connected layer. The latter is associated with the number of classes and serves as a classifier. The BNNeck block comprises two parts to obtain three different embeddings: the embeddings before the batch normalization, between the batch normalization and the classifier, and after the classifier. The first embedding is optimized for triplet loss, the second embedding is used for testing, and the third embedding is optimized for softmax loss.

### 2.3 Loss Function

Through the multi-branch network, the three-scale feature representations of global, local, and channel are obtained. However, due to the misalignment of non-global features, local features and channel features cannot use triplet loss. Inspired by the classification-metric loss function architecture in MGN, this paper uses softmax and triplet loss to construct the loss function.

**Softmax Loss.** The global feature, local feature, and channel feature passing through the BNNeck block are merged as the input feature of the loss function. For the merged feature  $f_i$ , the softmax loss function is:

$$L_{softmax} = - \sum_{i=1}^B \log \frac{e^{w_{y_i}^T f_i + b_{y_i}}}{\sum_{k=1}^C e^{w_k^T f_i + b_k}} \quad (1)$$

where  $B$  is the batch size,  $C$  is the number of classes of the current unlabeled training set  $X^t$ , the merged feature  $f_i$  is composed of  $\{\hat{g}, \hat{g}_{drop}, \hat{p}_g, \hat{p}_1, \hat{p}_2, \hat{c}_1, \hat{c}_2\}$  of the unlabeled sample  $x_i^t$ .

**Triplet Loss.** To further enhance the model performance, the batch-hard triplet loss [23] is used for all global features  $\{g, g_{drop}, p_g\}$  that do not pass through the BNNeck block. The batch-hard triplet loss is shown as follow:

$$L_{triplet} = - \sum_{i=1}^P \sum_{a=1}^K \left[ \alpha + \max_{p=1 \dots K} \|f_a^{(i)} - f_p^{(i)}\|_2 - \min_{\substack{n=1 \dots K \\ j=1 \dots P \\ j \neq i}} \|f_a^{(i)} - f_n^{(j)}\|_2 \right]_+ \quad (2)$$

where  $f_a^{(i)}$ ,  $f_n^{(i)}$ ,  $f_p^{(i)}$  are the merged features of anchor samples, positive samples, and negative samples, all of them are composed of  $\{g, g_{drop}, p_g\}$  from unlabeled samples  $x_i^t$ . The positive samples and the anchor samples have the same label, while the negative samples are the opposite. The positive sample farthest from the anchor sample and the negative sample nearest to the anchor sample constitute the triples of the batch-hard triplet loss. To further improve the robustness, use PK random sampling to select training samples, choose P classes randomly in each batch and randomly select K samples for each class. Positive samples, negative samples, and anchor samples are all in one batch.

The joint loss of comprehensive softmax and triplet loss is:

$$L = \lambda_1 L_{softmax} + \lambda_2 L_{triplet} \quad (3)$$

where  $\lambda_1$  and  $\lambda_2$  are hyperparameters that balance softmax loss and triplet loss.

## 2.4 Hierarchical Bottom-Up Clustering

When the training set has no labels, generating influential pseudo labels for the training set is a considerable challenge. At present, the mainstream method is to take the similarity and dissimilarity between samples as the supervision information, cluster the unlabeled data by the clustering algorithm, assign pseudo labels to unlabeled data according to the clustering results, and finally, train the network according to the generated pseudo labels. This paper uses the hierarchical clustering algorithm in BUC to create labels for unlabeled training samples.

**Cluster Preparation.** Since the training set  $X^t$  does not have labels, we assign different initial pseudo labels  $\{y_i = i | 1 \leq i \leq N\}$  to each training set sample. It is equivalent to treating every training sample as an individual class. The multi-branch network will maximize the distance between each sample during the first training. Then extract features of all training samples, calculate the similarity between sample pairs, merge the most similar samples into a cluster, and use the cluster-ID as the label.

**Cluster Merging.** After training on the training set assigned with different pseudo labels, the distance of the training sample pairs in the embedding space is maximized. Still, the similarity between samples of the same class is usually more significant than the similarity between samples of different classes. Based on this point, apply the hierarchical clustering algorithm to the multi-branch merging feature proposed above to merge the clusters from bottom to top. The quality of the pseudo labels used when training the model depends on the clustering results to integrate the same samples into the same cluster as possible and use the minimum distance criterion to calculate the similarity between clusters. Then the clusters with the most significant similarity are merged.

The minimum distance criterion takes the minimum distance of sample pairs in two clusters as the similarity. If two samples are in two clusters with high similarity, the two clusters are inclined to merge, no matter how different the other samples of the two clusters are. The advantage of this is that the same samples can be clustered together and assigned the same pseudo labels. The minimum distance criterion is described as follow:

$$D_{distance}(M, N) = \min_{x_m \in M, x_n \in N} d(x_m, x_n) \quad (4)$$

where  $M$  and  $N$  are two different clusters,  $d(x_m, x_n)$  is the Euclidean distance between any two samples of  $M$  and  $N$ .

In order to consider the speed and quality of clustering at the same time, the number of merging clusters  $m$  in each stage, the merging rate  $mp \in (0, 1)$ , and the number of merges  $N$  are specified. Where  $m = N \times mp$ . The number of clusters is  $C - N \times mp$  after  $t$  times of cluster merging.

**Diversity Regularization.** Similar samples will gradually cluster together in the clustering process, and the number of sample classes will decrease progressively. Assuming that the distribution of the training set is relatively uniform if the clustering result is correct, the samples with the same identity should be distributed in the same cluster. The distribution of the clusters should be relatively uniform, which means that each cluster should not contain too many samples. On the other hand, since the training samples have no ground truth labels, the clustering algorithm can easily merge similar but different samples into the same cluster and assign them the same pseudo labels, which reduces the clustering accuracy. In order to solve the above two problems, a diversity regularization term is proposed:

$$D_{diversity}(M, N) = |M| + |N| \quad (5)$$

where  $|M|$  represents the number of samples included in  $M$ .

The finally similarity calculation formula is:

$$D(M, N) = D_{distance}(M, N) + \alpha D_{diversity}(M, N) \quad (6)$$

where  $\alpha$  is a hyperparameter that balances similarity and diversity regularization, diversity regularization correlates with the number of samples in the cluster. The clustering algorithm tends to merge smaller clusters.

**Network Update.** In the first stage, use the pseudo labels generated by the clustering algorithm to update the multi-branch network. In the second stage, use the updated multi-branch network to extract the multi-branch features of all training sets. The similarity between the cluster is obtained according to the similarity calculation formula, and merge the most similar clusters. After clustering, assign the pseudo labels to the unlabeled training set according to the clustering result. The pseudo label is the ID of the cluster where the sample is located. Those two stages alternate until the model reaches the highest performance on the test set.

**Table.1** Introduction of Market-1501 and DukeMTMC-reID

		Market-1501	DukeMTMC-reID
Camera		6	8
Train	ID	751	702
	Num	12,936	16,522
Query	ID	750	702
	Num	3369	2228
Gallery	ID	750	702
	Num	19,732	17,661

### 3 Experiment Results

#### 3.1 Datasets and Protocols

We verify the proposed method on two benchmark re-ID datasets, such as Market-1501 and DukeMTMC-reID. Table 1 shows the basic parameters.

To evaluate the performance of the method proposed in this paper, the cumulative matching characteristics (CMC) at rank-1, rank-5, and rank-10, and mean average precision (mAP) are reported to measure the performance on the Market-1501 and DukeMTMC-reID comprehensively, and the re-rank is used to reorder the retrieval results.

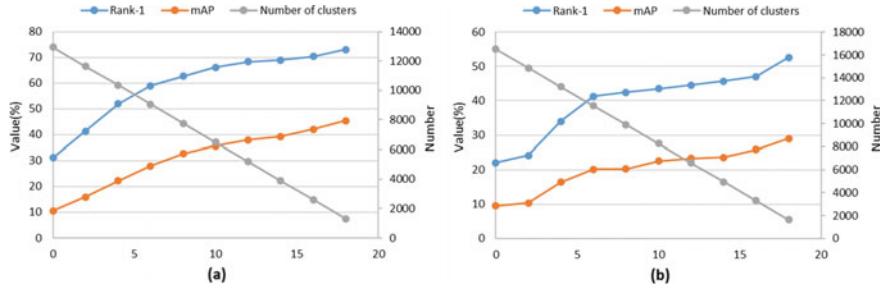
#### 3.2 Experimental Setting

**Train Details.** In the training phase, the inputs are adjusted to  $384 \times 128$ , and the data enhancement method is random cropping and random horizontal flipping. The batch size is set to 64. We adopt OSNet, which is pre-trained on the ImageNet [24], as the backbone network. Before the first clustering, the learning rate is initialized to 0.1, and after that, the learning rate is adjusted to 0.01. To balance the softmax loss and batch-hard triplet loss, both  $\lambda_1$  and  $\lambda_2$  in Eq. (3) are set to 1 and use the stochastic gradient descent (SGD) with a momentum of 0.9 to optimize the multi-branch network. In terms of clustering, the clustering speed  $mp$  is initialized to 0.05. we set the number of every training epoch to be 15 and obtain the highest performance at each stage. Each epoch is tested after the fifth epoch, saving the model with the highest performance and reading the model with the highest performance before extracting the multi-branch features required for the next clustering stage.

**Test Details.** In the testing phase, the input images are adjusted to  $384 \times 128$  and set the batch size to 32. Cosine distance is utilized to obtain the distance between the query samples and the gallery samples and use the re-ranking [25] based on k reciprocal nearest neighbor to optimize the distance matrix.

**Table.2** Comparison of the method in this paper with BUC

	Market-1501				DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
BUC	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5
Ours	71.7	83.7	87.6	50.9	52.9	66.0	71.4	35.1

**Fig. 3** Rank-1, mAP, and the number of clusters of each step

### 3.3 Compared with the Original Method BUC

We compose the method in this paper with the BUC based on the single-branch network on Market-1501 and DukeMTMC-reID. The comparison results are shown in Table 2. On the Market-1501, the method in this paper has achieved the highest performance of Rank-1 = 71.7% and mAP = 50.9%. Composed with the BUC based on the single-branch network, it has increased by 5.5 points and 12.6 points on Rank-1 and mAP. On the DukeMTMC-reID, the method in this paper has achieved the highest performance of Rank-1 = 52.9% and mAP = 35.1%. Composed with the BUC, it has increased by 5.5 points and 7.6 points on Rank-1 and mAP. The above results are optimized by re-ranking.

In Fig. 3, the changes in Rank-1 accuracy, mAP, and the number of clusters after each clustering. It can be seen that the closer the number of clusters is to the valid number of classes in the dataset, the better the performance of the model.

## 4 Conclusion

This paper applies a multi-branch network to replace the single-branch network in BUC and extracts global, local, and channel features simultaneously to obtain a better feature representation. Moreover, we also propose to utilize the Softmax-Triplet joint loss to optimize the model. After verification, the performance of the proposed method is better than that of the original BUC method.

**Acknowledgements** The paper was supported by the Natural Science Foundation of Shandong Province, China (Grant No. ZR2019MF073), the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) (Grant No. 20CX05001A), the Major Scientific and Technological Projects of CNPC (No. ZD2019-183-008), and the Creative Research Team of Young Scholars at Universities in Shandong Province (No. 2019KJN019).

## References

1. Zheng, L.: Person re-identification: past, present, and future. *Journal* (2016)
2. Kostinger, M.: Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288–2295 (2012)
3. Zhang, L.: Learning a discriminative null space for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1239–1248 (2016)
4. Zheng, L.: Scalable person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124 (2015)
5. Zheng, Z.: Joint discriminative and generative learning for person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2133–2142 (2019)
6. Chang, X.: Multi-level factorisation net for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2109–2118 (2018)
7. Sun, Y.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). *Journal* (2017)
8. Ploco, A.: Spatial-temporal omni-scale feature learning for person re-identification. In: 2020 8th International Workshop on Biometrics and Forensics (IWBF)
9. Wang, G.: Spatial-temporal person re-identification. *Journal* (2018)
10. Wang, D.: Unsupervised person re-identification via multi-label classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10978–10987 (2020)
11. Lin, Y.: Unsupervised person re-identification via softened similarity learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3387–3396 (2020)
12. Zeng, K.: Hierarchical clustering with hard-batch triplet loss for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13654–13662 (2020)
13. Yutian, L.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI2019 (2019)
14. Wang, J.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2275–2284 (2018)
15. Liu, C.: Unity style transfer for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6886–6895 (2020)
16. Yang, Q.: Patch-based discriminative feature learning for unsupervised person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3628–3637 (2019)
17. Yang, F.: Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. *Journal* (2021)
18. Herzog, F.: Lightweight multi-branch network for person re-identification. *Journal* (2021)
19. Guangshuo, W.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)

20. Hao, C.: Learning discriminative and generalizable representations by spatial-channel partition for person re-identification. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 2483–2492 (2020)
21. Kaiyang, Z.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3702–3712 (2017)
22. Hao, L.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
23. Hermans, A.: In defense of the triplet Loss for person re-Identification. Journal (2017)
24. Krizhevsky, A.: Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105 (2012)
25. Zhong, Z.: Re-ranking person re-identification with k-reciprocal encoding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3652–3661 (2017)

# Affine Non-negative Hybrid Collaborative Representation Based Classification



Haoquan Guan Baodi Liu Weifeng Liu Kai Zhang Ye Li and Peng Liu

**Abstract** Representation based classification methods have achieved essential results in image classification recent in recent years. Primarily, the proposed affine non-negative representation based classification (ANCR) model has achieved satisfactory results, but ANCR does not consider the specific class representation for a test sample. We offer an affine non-negative hybrid collaborative representation based classification (ANHC) algorithm for image classification. Specifically, the ANHC method combines the traditional representation with specific class of collaborative representation to represent test samples. Then, we introduce affine transformation to use the image features of the affine subspace to obtain better classification performance. Various experiments on face recognition data and handwriting recognition datasets show that the ANHC method is superior to several traditional methods in classification accuracy.

**Keywords** Non-negative representation · Image classification · Collaborative representation · Specific class

---

H. Guan

College of Oceanography and Space Informatics, China University of Petroleum (East China),  
Qingdao, China

B. Liu · W. Liu

College of Control Science and Engineering, China University of Petroleum (East China),  
Qingdao, China

e-mail: [liubaodi@upc.edu.cn](mailto:liubaodi@upc.edu.cn)

P. Liu

Shandong Kexun Information Technology Co., Ltd., Qingdao, China

Y. Li

Qilu University of Technology Shandong Academy of Sciences, Jinan, China

K. Zhang

School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China

## 1 Introduction

Image classification has received widespread attention in the application of face recognition and other practical problems in recent years. The representation based classifier model has aroused extensive research interest in the field of visual recognition.

Traditionally, the nearest neighbor (NN) [1] method and nearest subspace (NS) [2] method are widely used in image classification. Then, Wright [3] proposed sparse classical representation based classification (SRC), SRC uses all training samples to sparsely represent test samples by solving the  $L_1$ -regularized minimization problem and then queries which class produces the smallest residual of the test sample to classify test sample. Still, this kind of processing method increases the time-consuming to solve the encoding vector. To solve this problem, Zhang [4] proposed a collaborative representation based classification (CRC) method that uses the  $L_2$ -norm as the regularization term, dramatically shortening calculation time. Based on the traditional sparse representation (SRC) and collaborative representation (CRC) models. Liu [5] proposed a local linear K-nearest neighbor method. Lai [6] proposed a sparse representation method to obtain the optimal representation of test samples by minimizing the sparseness of training samples. Shao [7] used synthetic faces to optimize the comprehensive dictionary dynamically and proposed a new classification algorithm. Akhtar [8] proposed a sparse augmented collaborative representation method (SA-CRC), which uses sparse representation to enhance dense collaborative representation. Li [9] designed a sparse enhancement weighted collaborative representation image classification method. Zheng [10] developed a collaborative representation classification method based on k-nearest classes. Gou [11] designed a weighted discriminant collaborative representation classification method. Xu [12] There are negative element codes in the classification methods of CRC, SRC, and their variants, which leads to errors in sample classification. Under the enlightenment of non-negative matrix factorization (NMF) [13], they designed a non-negative representation classification method (NRC). Zhao [14] proposed a Laplacian-regularized non-negative representation algorithm, which mainly used for task of clustering and dimensionality reduction. Yin [15] developed a class specific residual constraint non-negative representation, which is used for pattern recognition. Benuma [16] proposed a sparse representation that is sensitive to the position of the kernel, which is an algorithm for face recognition. Yin [17] proposed affine non-negative collaborative representation based classification (ANCR), various experiments have proved that the ANCR method has obvious advantages over the representation based classification method. However, in the ANCR model, the training samples of all classes represent the test samples, and the specific class of the test sample is not considered. We propose an affine non-negative hybrid collaborative representation based classification (ANHC) model. Specifically, the ANHC model combines the traditional class with specific class of collaborative representation to represent test samples and introduces affine non-negative constraint to use the image feature of the affine subspace.

The main contributions of this paper are as follows,

- We propose an affine non-negative hybrid collaborative representation-based classification (ANHC) model, which combines the traditional class with specific class collaborative representation.
- We introduce affine non-negative constraint to use the image feature of the affine subspace to restrict features in aligned subspace.
- We conduct experimental verification on four common datasets, and prove the effectiveness of the ANHC model through experimental data analysis.

## 2 Related work

### 2.1 Sparse Representation Based Classification (SRC)

Wright [3] designed sparse representation based classification models. SRC directly uses all training samples as dictionaries, obtains a sparse matrix by solving the L<sub>1</sub>-regularized minimization problem, and then matches the test samples by calculating the minimum reconstruction error. Suppose  $n$  training samples belong to class C, and the training data matrix is  $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^C]$ , where  $\mathbf{X}^i$  is the data matrix of the  $i$ -th class. The  $i$ -th class has  $n_i$  training samples,  $d$  is the dimension of the vectorized sample. For the test sample  $\mathbf{y}$ , it needs to be classified by solving the sparse representation coefficient  $\mathbf{s}$ , that is, the vector  $\mathbf{s}$  can be solved by Formula (1).

$$\arg \min_{\mathbf{s}} \|\mathbf{y} - \mathbf{X}\mathbf{s}\|_2^2 + \beta \|\mathbf{s}\|_1 \quad (1)$$

After obtaining the sparse representation, the label of  $\mathbf{y}$  can be obtained by calculating the smallest residual.

$$id(\mathbf{y}) = \arg \min_i \|\mathbf{y} - \mathbf{X}^i \mathbf{s}^i\|_2^2 \quad (2)$$

where  $id(\mathbf{y})$  is the label of  $\mathbf{y}$ .

### 2.2 Collaborative Representation Based Classification (CRC)

Unlike SRC, the CRC model proposed by Zhang [4] plays a crucial role in collaborative representation and uses L<sub>2</sub>-norm in SRC. The final expression of CRC is shown in formula (3).

$$\arg \min_{\mathbf{s}} \|\mathbf{y} - \mathbf{X}\mathbf{s}\|_2^2 + \beta \|\mathbf{s}\|_2^2 \quad (3)$$

The parameter  $\beta$  is extremely significant for adjusting the cooperative representation. After obtaining the sparse representation, the label of  $\mathbf{y}$  can be obtained by calculating the smallest residual.

$$id(\mathbf{y}) = \arg \min_i \|\mathbf{y} - \mathbf{X}^i \mathbf{s}^i\|_2^2 \quad (4)$$

### 2.3 Non-negative Representation Based Classification (NRC)

The core idea of CRC and SRC is to encode the test sample  $\mathbf{y}$  on the entire training sample matrix  $\mathbf{X}$ . However, CRC and SRC are prone to produce negative elements and then rebuild the sample by adding and subtracting the training sample, which is prone to misclassification. Based on the above problems, Xu [12] imposed non-negative constraints on the encoding coefficients and not use L<sub>1</sub>-norm or L<sub>2</sub>-norm regularization in the target formula. The target formula uses a model based on non-negative representation to calculate the encoding vector:

$$\arg \min_s \|\mathbf{y} - \mathbf{X}s\|_2^2 \text{ s.t. } s \geq 0 \quad (5)$$

The classification process of NRC is similar to SRC and CRC.

### 2.4 Affine Non-negative Collaborative Representation Based Classification (ANCR)

Although the NRC method has achieved excellent results in classification tasks, there are also two main shortcomings in NRC. First, The NRC discards the regularization term, which could cause misclassification of the results. Second, the NRC ignores image features hidden in the affine subspace. Regarding the issue above, Yin [17] proposed the ANCR method. ANCR introduces regularization constraints on the basis of the NCR target formula. In addition, ANCR introduces affine constraints to search the image features in the affine subspace. The final expression of ANCR is shown in Formula (6).

$$\arg \min_s \|\mathbf{y} - \mathbf{X}s\|_2^2 + \beta \|s\|_2^2, \text{ s.t. } s \geq 0, \mathbf{1}^T s = 1 \quad (6)$$

The classification process of ANCR is similar to SRC and CRC.

### 3 Proposed Method

#### 3.1 *Affine Non-negative Hybrid Collaborative Representation Based Classification (ANHC)*

To overcome the shortcomings of ANCR, we propose to combine the collaborative representation of traditional and specific classes and introduce affine non-negative constraints to align the image features hidden in the affine subspace. In summary, the final expression of ANHC is shown in Formula (7).

$$\arg \min_s \|y - Xs\|_2^2 + \lambda \sum_{i=1}^C \left\{ \|y - X^i s^i\|_2^2 + \beta \|s^i\|_2^2 \right\}, \text{ s.t. } s \geq 0, 1^T s = 1 \quad (7)$$

The first item is the reconstruction error term, which ensures that each class can participate in the classification. The second item is a collaborative representation of a specific class used to ensure that different training classes have unique contributions to classification. The last constraint is the non-negative affine constraint item, and they explore the image features hidden in the affine subspace and ensure the non-negativity of the encoding vector.

#### 3.2 *Optimization*

For the convenience of calculation, we temporarily ignore the constraints and rewrite the formula into the following form:

$$\begin{aligned} f(s) &= y^T y - 2y^T Xs + s^T X^T Xs + \beta s^T s \\ &\quad + \lambda Cy^T y - 2\lambda y^T Xs + \lambda \sum_{i=1}^C \{s^{iT} X^{iT} X^i s^i\} \end{aligned} \quad (8)$$

In Eq. (8), the specific class can be simplified.

$$\begin{aligned}
& \sum_{i=1}^C \{\mathbf{s}^{iT} \mathbf{X}^{iT} \mathbf{X}^i \mathbf{s}^i\} \\
& = \lambda \mathbf{s}^T \sum_{i=1}^C \{\mathbf{X}^{iT} \mathbf{X}^i\} \mathbf{s} \\
& = \lambda \mathbf{s}^T \begin{pmatrix} \mathbf{X}^{1T} \mathbf{X}^1 & & \\ & \ddots & \\ & & \mathbf{X}^{CT} \mathbf{X}^C \end{pmatrix} \mathbf{s} \\
& = \lambda \mathbf{s}^T \mathbf{A} \mathbf{s}
\end{aligned} \tag{9}$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{X}^{1T} \mathbf{X}^1 & & \\ & \ddots & \\ & & \mathbf{X}^{CT} \mathbf{X}^C \end{pmatrix} \tag{10}$$

So, the final expression of (8) is rewritten in Formula (11).

$$f(s) = (1 + \lambda C) \mathbf{y}^T \mathbf{y} - 2(1 + \lambda) \mathbf{y}^T \mathbf{X} \mathbf{s} + s^T (\lambda \mathbf{A} + \mathbf{X}^T \mathbf{X} + \beta \mathbf{I}) \mathbf{s} \tag{11}$$

Then, we reconsider the influence of non-negative constraints and affine constraints on optimization, introduce the variable  $\mathbf{z}$ , and use the ADMM algorithm [18] to optimize the formula (11). The optimized expression of (11) is shown in formula (12).

$$\begin{aligned}
f(s, z) & = (1 + \lambda C) \mathbf{y}^T \mathbf{y} - 2(1 + \lambda) \mathbf{y}^T \mathbf{X} \mathbf{s} \\
& + s^T (\lambda \mathbf{A} + \mathbf{X}^T \mathbf{X} + \beta \mathbf{I}) \mathbf{s} \text{ s.t. } \mathbf{s} = \mathbf{z}, \mathbf{z} \geq 0, \mathbf{1}^T \mathbf{z} = 1
\end{aligned} \tag{12}$$

The Lagrangian function of (12) is:

$$\begin{aligned}
L_p(c, z, \delta, \rho) & = (1 + \lambda C) \mathbf{y}^T \mathbf{y} - 2(1 + \lambda) \mathbf{y}^T \mathbf{X} \mathbf{s} \\
& + s^T (\lambda \mathbf{A} + \mathbf{X}^T \mathbf{X} + \beta \mathbf{I}) \\
& + \langle \delta, \mathbf{z} - c \rangle + \frac{\rho}{2} \|\mathbf{z} - c\|_2^2
\end{aligned} \tag{13}$$

Let  $\frac{\partial L_p(s)}{\partial s} = 0$ , with fixed  $\delta$  and  $\mathbf{z}$ , so  $\mathbf{s}$  can be obtained:

$$s_{t+1} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{A} + \beta \mathbf{I} + \frac{\rho}{2} \mathbf{I})^{-1} [(1 + \lambda) \mathbf{X}^T \mathbf{y} + \frac{\rho \mathbf{z} + \delta}{2}] \tag{14}$$

To update  $\mathbf{z}$ , we fixed  $\mathbf{s}$  and  $\delta$ .

$$\mathbf{z}_{t+1} = \min_z \left\| \mathbf{z} - \mathbf{s} + \frac{\delta}{\rho} \right\|_2^2, \text{ s.t. } \mathbf{z} \geq 0, \mathbf{1}^T \mathbf{z} = 1 \quad (15)$$

Huang [19] has solved the proof and solution process of the above formula (15). To update  $\delta$ , we fixed  $\mathbf{s}$  and  $\mathbf{z}$ .

$$\delta_{t+1} = \delta_t + \rho(z_{t+1} - c_{t+1}) \quad (16)$$

### 3.3 Classification

Give a test sample  $\mathbf{y} \in \mathbb{R}^{d \times 1}$ , the encoding vector  $\mathbf{s}$  can be obtained by Formula (14), and then the test sample is assigned to the class with the smallest reconstruction error as shown in the following.

$$id(\mathbf{y}) = \arg \min_i \|\mathbf{y} - \mathbf{X}^i \mathbf{s}^i\|_2^2 \quad (17)$$

The classification process of ANHC is similar to the previous models, and main procedure of the ANHC model is as follows.

**Algorithm 1** ANHC algorithm.

**Require:** Training matrix  $\mathbf{X}$ , episode size T, parameter  $\lambda, \beta, \rho \geq 0$ , and test sample of  $\mathbf{y}$ .

**Ensure:** Identity of  $\mathbf{y}$

- 1: Initialize
- 2: **for**  $t = 1, 2, \dots, T$  **do**
- 3:     Calculate  $\mathbf{c}$  by (14);
- 4:     Calculate  $\mathbf{z}$  by (15);
- 5:     Calculate  $\delta$  by (16);
- 6:     Calcuate  $f(T)$  according to (8);
- 7: **end for**
- 8: **for**  $i = 1; i \leq 0; i++$  **do**
- 9:      $w^i(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}^i \mathbf{s}^i\|_2^2$
- 10: **end for**
- 11:  $id(\mathbf{y}) = \arg \min_s (w^i)$ ;
- 12: return  $id(\mathbf{y})$

**Table 1** Classification accuracy(%) of several models on the AR dataset

Method	54	120	200
SVM	81.6	89.3	91.6
CRC	78.7	88.1	91.0
SRC	82.1	88.3	90.3
ProCRC	81.4	90.7	93.7
NRC	85.2	91.3	93.3
ANCR	85.7	91.3	94.2
<b>ANHC</b>	<b>86.0</b>	<b>91.8</b>	<b>94.4</b>

## 4 Experiment

We verify the proposed ANHC model on four benchmark datasets, such as AR, CMU PIE, USPS, and MNIST datasets. The following subsection will illustrate the experimental results in detail.

### 4.1 Experiments on the AR Dataset

The AR dataset [20] contains face images of 126 individuals, with a total of more than 4000 face images, including 26 frontal images. In the experimental setting, all images are adjusted to  $60 \times 43$  pixels. These images are facial data from 50 men and 50 women. All the image is divided into 100 class. Each class selects seven images as training samples and test samples. They are projected into 54, 120, and 300-dimensional subspaces through principal component analysis (PCA). The experimental results of the ANHC model and several other models are shown in Table 1. From Table 1, we can see that the ANHC model has achieved 86, 91.8, 94.4% classification accuracy in different dimensions. Under the same experimental conditions, the accuracy of ANHC is 0.3%, 0.5%, and 0.2% higher than that of ANCR, respectively.

### 4.2 Experiments on the CMU PIE Dataset

The CMU PIE dataset [21] has gradually become an important test set for face recognition. This is a dataset of facial poses, lights, and facial expressions composed of more than 4000 different facial photos of 68 volunteers. There are a total of 68 other classes of objects. The CMU PIE dataset has a capacity of 11,554 images, and all the images are cropped to  $32 \times 32$  pixels. In the experiment, we select 50 images in each class as training samples. The experimental results of the ANHC model and several other models are shown in Table 2. From Table 2, the accuracy

**Table 2** Classification accuracy(%) of several models on the CMU PIE dataset

Method	CMU PIE
SVM	82.6
CRC	86.3
SRC	87.6
ProCRC	89.4
NRC	90.2
ANCR	90.4
<b>ANHC</b>	<b>91.4</b>

of ANHC is 91.4%, which is 1% and 1.2% higher than the accuracy of ANCR and NRC, respectively. The accuracy of ANHC model is 1% and 1.2% higher than that of ANCR and NRC respectively.

### 4.3 Experiments on the USPS Dataset

USPS dataset [22], there are about 7291 handwritten images for training examples, and 2007 images for test samples. The pixel size of the image is  $28 \times 28$ , the value of the number is 0–9. We selected training samples that are changed according to the value of  $N$ , the  $N$  is the number of training samples selected in each class. With the increase of training images of each class, the recognition accuracy of ANHC increases steadily. The experimental results of the ANHC model and several other models are shown in Table 3. Table 3 clearly shows the accuracy of ANHC in the case of different numbers of training samples, reaching the accuracy of 94.4%, 95.2%, and 96.0% respectively. The ANHC reaches the highest accuracy when  $N$  is 200. Compared with the ANCR method, when the number of samples is 50, the ANHC model improves by 2.8%. However, when the number of samples increases to 200, it only increases by 0.8%. This shows that the ANHC model has a more obvious improvement effect when there are fewer training samples.

**Table 3** Classification accuracy(%) of several models on the USPS dataset

Method	50	100	200	300
SVM	90.1	91.6	93.8	94.0
CRC	87.8	89.7	90.9	91.6
SRC	87.1	88.6	89.8	90.5
ProCRC	90.9	91.9	92.2	92.2
NRC	90.3	91.6	92.7	93.0
ANCR	91.6	93.6	94.7	95.2
<b>ANHC</b>	<b>94.4</b>	<b>95.2</b>	<b>95.4</b>	<b>96.0</b>

**Table 4** Classification accuracy (%) of several models on the MNIST dataset

Method	50	100	200	300
SVM	80.6	81.0	82.5	82.6
CRC	80.7	84.1	86.6	86.7
SRC	81.0	85.4	86.1	86.9
ProCRC	89.8	91.4	91.6	91.7
NRC	90.1	91.2	92.3	93.0
ANCR	87.6	90.8	93.7	94.2
ANHC	<b>91.1</b>	<b>92.4</b>	<b>93.8</b>	<b>94.3</b>

#### 4.4 Experiments on the MNIST Dataset

MNIST dataset [23], there are about 6000 handwritten images for training examples, and 10,000 images for test samples. These numbers have been standardized in size and are located in the center of the image. The pixel size of the image is  $28 \times 28$ , the value of the number is 0–9. In the experiment,  $N$  images are selected as training samples in each type of sample. The classification results of ANHC and several other models are recorded in Table 4. In Table 4, the classification accuracy of ANHC has achieved different classification results in four training matrices with different values, reaching accuracy of 91.1%, 92.4%, 93.8%, and 94% respectively. In the process of increasing the number of training samples from 50 to 200, the classification effect of ANHC becomes more significant.

#### 4.5 Parameter Sensitiveness Analysis

Boyd [18] proved the convergence of bivariate ADMM, which is also applicable to ANHC method. As the iterations increase, the objective function gradually decreases. It is empirically proved that the proposed method is convergent.

In the experiments, we perform the proposed ANHC method with different combinations of two parameters,  $\beta$ , and  $\lambda$ .  $B$  is an essential parameter in the ANHC algorithm, which is used to adjust the tradeoff between the reconstruction error and the specific class collaborative representation.  $\Lambda$  is also a necessary parameter in the model, which is used to control the tradeoff between the traditional specific class collaborative representation and the specific class collaborative representation. In the experiment, we performed experiments on four data sets. Selecting the AR data as an example, the value of  $\beta$  is set to [0.0001,0.01]. When the value of  $\beta$  increases from 0.01 to 0.1, the model's performance will also decrease. The value range of the value of  $\lambda$  is the same as the value of  $\beta$ . When the value of  $\lambda$  increases, the model's performance will also decrease, because  $\lambda$  is too large, which enhances the model more inclined to a specific class and ignores the cooperative representation

mechanism. In summary, to enhance the classification performance, we set  $\beta$  and  $\lambda$  to 0.0001 and 0.001, respectively.

## 4.6 Ablation Study

In this subsection, we discard all constraint terms and specific class term to discuss the influence of different terms on the experimental results.

The first model can be obtained by discarding the specific class item and affine non-negative constraints term in (7). Actually, the first model is the CRC model.

The second model can be obtained by removing the affine non-negative constraint term in (7),

$$\arg \min_s \|y - Xs\|_2^2 + \beta \|s\|_2^2 + \lambda \sum_{i=1}^C \|y - X^i s^i\|_2^2 \quad (18)$$

We named the second baseline the specific class collaborative representation (SCR) model. Through mathematical operations, the solution of the above SCR is shown in Formula (19).

$$s = (1 + \lambda)(X^T X + \beta I + \lambda A)^{-1} X^T y \quad (19)$$

The third model can be obtained by discarding the specific class item in (7). Actually, the third model is the ANCR model.

Table 5 summarizes ANHC and the above three models. Experiments are performed on four datasets, and the experimental results of the ANHC model and several other models are shown in Table 6. According to Table 6, we can obtain the following results:

- (1) The performance of SCR is better than that of CRC, which shows that the specific class items enrich the training samples so that the code vector of SCR contains more identification information and improves the accuracy of SCR.

**Table 5** Summary of ANHC and the three baseline models

Mothed	Regularization terms	Constrains	
		$s \geq 0$	$\sum_{i=1}^C \ y - X^i s^i\ _2^2$
CRC	$\min_s \ y - Xs\ _2^2 + \beta \ s\ _2^2$	✗	✗
SCR		✗	✓
ANCR		✓	✗
ANHC		✓	✓

**Table 6** Recognition accuracy(%) of four models in AR dataset

Method	54	120	200
CRC	78.7	88.1	91.0
SCR	84.7	90.4	93.7
ANCR	85.7	91.3	94.2
ANHC	<b>86.0</b>	<b>91.7</b>	<b>94.4</b>

- (2) The performance of ANCR is better than CRC, which shows that non-negative constraints suppress negative elements in the encoding vector, and affine constraints use image features in the affine subspace to enhance the classification performance of ANCR.
- (3) The classification effect of the ANHC model is more obvious than the other three baseline models, which shows that the ANHC method is effective.

## 5 Conclusion

This paper proposes an affine non-negative hybrid collaborative based classification (ANHC) algorithm for image classification to exert both traditional class and specific class collaborative representation into complexities. Moreover, we introduce affine non-negative constraints to cope with affine subspace structure hidden in image features to enhance classification performance further. Extensive experiments on four image datasets have proved the superiority of affine non-negative hybrid collaborative representation based classification.

**Acknowledgements** The paper was supported by the Natural Science Foundation of Shandong Province, China (Grant No. ZR2019MF073), the Open Research Fund from Shandong Provincial Key Laboratory of Computer Network (No. SDKLCN-2018-01), the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) (Grant No. 20CX05001A), the Major Scientific and Technological Projects of CNPC (No. ZD2019-183-008), and the Creative Research Team of Young Scholars at Universities in Shandong Province (No. 2019KJN019).

## References

1. Jadbabaie, A., Jie, L., Morse, A.S.: Coordination of groups of mobile autonomous agents using nearest neighbor rules. In: Proceedings of the 41st IEEE Conference on Decision and Control, IEEE Transactions on Automatic Control, vol. 48, pp. 988–1001. IEEE (2003)
2. Jentzung, C., Chiachen, W.: Discriminant waveletfaces and nearest feature classifiers for face recognition **24**(12), 1644–1649 (2002)
3. Wright, J., Allen, Y.: Ganesh A. Robust face recognition via sparse representation. **31**(2), 210–227 (2009)
4. Lei, Z., Meng, Y., Xiangchu, F.: Sparse representation or collaborative representation: WhicHelps face recognition. In: IEEE International Conference on Computer Vision, pp. 471–478. IEEE, Barcelona Spain (2011)

5. Qingfeng, L., Chengjun, L.: A Novel locally linear KNN method with applications to visual recognition. 28(9), 2010–2021 (2017)
6. Jian, L., Xudong, J.: Class-wise sparse and collaborative patch representation for face recognition 25(7), pp. 3261–3272 (2016).
7. Changbin, S., Xiaoning, S., Zhenhua F.: Dynamic dictionary optimization for sparse-representation-based face classification using local difference images 393, 1–14 (2017)
8. Akhtar, N., Shafait, F., Mian, A.: Efficient classification with sparsity augmented collaborative representation **65**, 136–145 (2017)
9. Ziqi, L., Jun, S., Xiaojun, W., Hefeng, Y.: Sparsity augmented weighted collaborative representation for image classification. 28(5), 053032 (2019)
10. Chengyong, Z., Ningning, W.: Collaborative representation with k-nearest classes for classification **117**, 30–36 (2018)
11. Jianping, G., Lei, W., Zhang, Y., Yunhao, Y., Weihua, A.: Weighted discriminative collaborative competitive representation for robust image classification **125**, 104–120 (2020)
12. Jun, X., Wangpeng, A., Lei, Z., David, Z.: Sparse, collaborative, or nonnegative representation: which helps pattern classification. Pattern Recogn. **88**, 679–688 (2018)
13. Daniel, D.L., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. 401(6755), 788–791 (1999)
14. Yinping, Z., Long, C.: Laplacian regularized non-negative representation for clustering and dimensionality reduction. In: IEEE Transactions on Circuits and Systems for Video Technology (2020)
15. Hefeng, Y., Xiaojun, W.: Class-specific residual constraint non-negative representation for pattern classification 29(2), 023014 (2019)
16. Benuwa, B., Ghansah, B., Ansah, E.K.: Kernel based locality sensitive discriminative sparse representation for face recognition. In: Scientific African, vol. 7, pp. e00249 (2020)
17. Hefeng, Y., Xiaojun, W., Zhenhua, F., Josef, K.: Affine non-negative collaborative representation based pattern classification. [arXiv:2007.05175](https://arxiv.org/abs/2007.05175) (2020).
18. Boyd, S., Parikh, N.: Distributed optimization and statistical learning via the alternating direction method of multipliers 3(1), 1–122 (2010)
19. Huang, J., Nie, F., Huang, H.: A new simplex sparse learning model to measure data similarity for clustering. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
20. Martinez, A. M.: The ar face database. CVCTechnical Report 24 (1998)
21. Sim, T., Baker, S., Bsat, M.: The CMU Pose, Illumination, and Expression(PIE) database. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp.53–58. IEEE Computer Society (2002)
22. Hull, J.J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. 16(5), 550–554 (1994)
23. Lecun, Y., Bottou, L.: Gradient-based learning applied to document recognition **86**(11), 2278–2324 (1998)

# Pathologist-Level Classification of Melanoma Disease Pathologies Using a Convolutional Neural Network: A Retrospective Study of Chinese



Tao Li , Fangfang Li , Jie liu , and Ke Zuo

**Abstract** Melanoma is a highly malignant skin tumor which causes nearly half of skin cancer deaths. However, in China, especially in rural area, professional pathologists who can diagnose melanoma early and correctly are insufficient. Therefore, there is a need to develop an objective and quantitative method for melanoma diagnosis. The purpose of this study is to illustrate the potential of deep learning to assist pathologists in assessing the histopathological melanoma diagnosis of Chinese patients. We established a novel convolutional neural network model for melanoma diagnosis. The work was carried out on a histopathology database with 633 digital whole-slide images from 314 patients in China. Then, the model achieved pathologist-level classification of the pathological images, with an accuracy of 0.92. These findings suggest that convolutional neural networks can be an efficient tool to assist pathologists to diagnose melanoma in Chinese patients, with low time cost and high accuracy.

**Keywords** Melanoma · Histopathology · Deep learning · Precision medicine · Image analysis

## 1 Introduction

Malignant melanoma is a melanoma cell carcinoma [1, 2]. According to the Global Cancer Statistic, over 60,000 patients with melanoma die from the disease each year, while another 280,000 new cases are diagnosed [1]. As for now, the pathologist's accurate diagnosis of hematoxylin and eosin (H&E) stained tissue slides is the key

---

T. Li · J. liu · K. Zuo ()

National University of Defense Technology, Changsha 410073, Hunan, China  
e-mail: [zuko@nudt.edu.cn](mailto:zuko@nudt.edu.cn)

F. Li

The Department of Dermatology, Xiangya Hospital, Central South University, Changsha 410008, Hunan, China

Hunan Key Laboratory of Skin Cancer and Psoriasis, Changsha 410008, Hunan, China

Hunan Engineering Research Center of Skin Health and Disease, Changsha 410008, Hunan, China

to the diagnosis and successful treatment planning for melanoma [3–8]. However, the pathologist and people ratio is as large as 1:100,000 in China. In addition, most experienced pathologists are located in “AAA” hospital of major cities, which has further exacerbated the scarcity of pathologists in the rural China. Therefore, many patients are unaware of having melanoma and may not consult in a tertiary hospital for further help. Therefore, there is a need to develop an objective and quantitative automatic diagnosis method for melanoma diagnosis.

Previous studies on histopathology melanoma whole slide images (WSIs) mainly used computer-based image analysis approaches for cell segmentation [4], invasion depth prediction [5], et al. These works are based on topology, statistic, or machine learning, etc. However, due to the technological limitations, the high performance of these works was confined to the small handpicked dataset, which limits its clinical application.

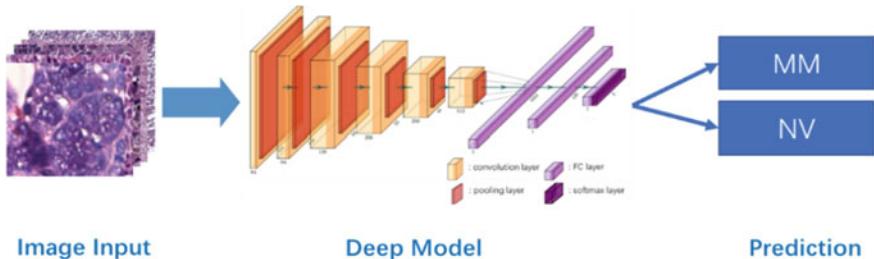
The way has changed in recent years, deep learning has deeply affected medical image analysis. Diagnostic convolutional neural networks (CNN) can match or exceed the ability of field experts in pathological image recognition tasks [9], including the diagnosis of lung cancer [10] and breast cancer recognition [11]. In melanoma pathology recognition task, Achim Hekler et al. demonstrated pathologist-level classification of malignant melanomas versus benign nevi with a pretrained ResNet50 CNN [12]. In the eyelid malignant melanoma identification task, the study based on CNN and random forest obtained an area under curve (AUC) of 0.998 on 155 eyelid WSIs [13]. Kulkarni et al. proposed a deep learning based method for disease-specific survival prediction in early stage melanoma and achieved a 0.905 AUC [14]. However, skin color is an important basis for melanoma diagnosis. Current researches have focused on the white or black skin, with little research on the yellow skin.

This study is the first work to apply deep learning to the diagnosis of pathologic melanoma in Chinese patients. The purpose of this study is to illustrate the potential of deep learning in assisting pathologist in the diagnosis of melanoma in Chinese patients. A novel CNN model was built for melanoma diagnosis. And we established a Chinese histopathology image database of 633 WSIs from 314 patients for model training and evaluation. As the result shows, the model achieves an accuracy of 0.92 and shows a strong potential in Chinese melanoma diagnosis, which achieved a diagnostic efficacy comparable to that of pathologist.

## 2 Method

### 2.1 Model

CNN is a special multilayer neural network that recognizes complex visual patterns extracted from simple preprocessed pixel images [9]. To achieve a balance between model performance and efficiency, we build a CNN model that is simple enough for



**Fig. 1** The models identify two disease types in the image tiles. Six convolutional layers, six pooling layers and three dense layers are contained in the model

doctors to understand, we elaborately designed a CNN model. As shown in Fig. 1, the model designed in this paper contains six convolution layers, three fully connected layers, and one softmax layer. The image features were extracted by convolutional layers, and classified by the fully connected layers.

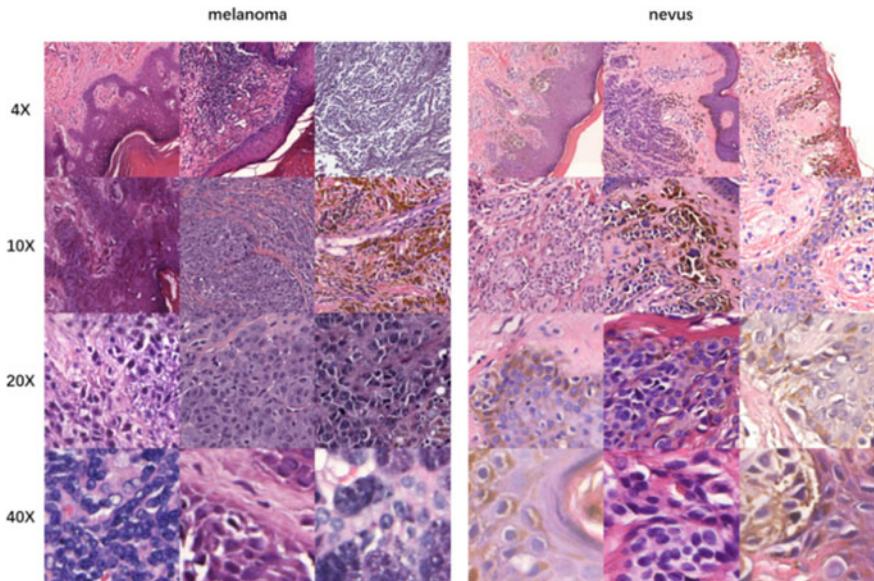
## 2.2 Dataset

This study was performed under the Declaration of Helsinki Principles [15]. Collaborating with Central South University Xiangya Hospital (CSUXH), we collected 633 H&E stained whole-slide histopathology images and built a multicenter pathological image database from March 2014 to May 2019. 184 melanoma WSIs and 449 nevus WSIs were included. The WSIs were scanned by the 3Dhistech PANNORAMIC MIDI scanner. All the patients are Chinese.

In the study, WSIs of melanoma and three common skin diseases including compound nevi, junctional nevi, and intradermal nevi were collected. And all WSIs are clear enough for diagnosis. All of the images were labeled by three pathologists and reviewed by two experts.

H&E stained WSIs were acquired at magnifications of  $\times 4$  to  $\times 40$  by scanner, with 10,000 to over 100,000 pixels in each dimension. It can be challenging to use CNN for visual analysis in an exhaustive way. To solve this problem, all the WSIs were cut into  $256 \times 256$  pixel patches at four different magnifications:  $4\times$ ,  $10\times$ ,  $20\times$  and  $40\times$ , respectively. And the background patches were removed by OSTU method. Figure 2 shows patch examples of each disease at different magnifications.

To ensure that the model could be trained and validated efficiently, we randomly selected 20,000 patches from four magnifications respectively. The patches in four magnifications were divided into training set, validation set and test set, with a patch ratio of 7:1.5:1.5. And the data of one patient will only be divided into one of three sets to ensure no cross-contamination of data. As a result, 56,000 patches were generated in the training set, 12,000 in the validation set and 12,000 in the test set.



**Fig. 2** Sample patches from the dataset at different magnifications. Melanoma and nevus patches are shown. Patches at 4 $\times$ , 10 $\times$ , 20 $\times$ , and 40 $\times$  magnifications show different morphological features

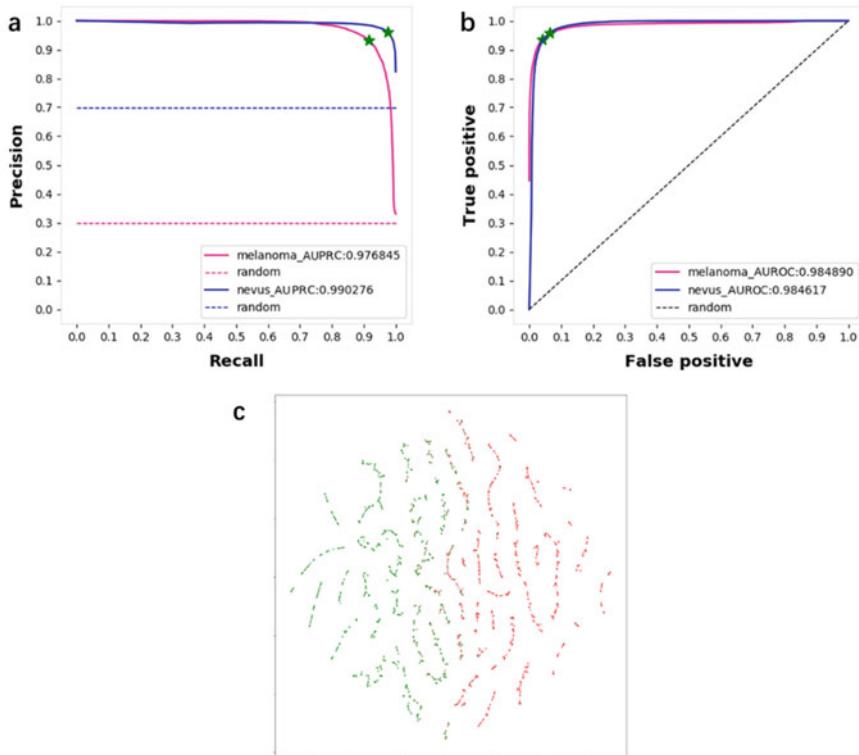
### 2.3 Training Progress

In the model training progress, we used cross-entropy loss and stochastic gradient descent (SGD) optimization, with a learning rate of 0.01, the momentum of 0.9, and the weight decay of 0.0001. The model was trained in a single TITAN RTX GPU.

## 3 Result

The performance of the model is shown in Fig. 3. When melanoma was used as a positive sample of the test set, the accuracy of the model was 92.5%, and its Area Under ROC (AUROC) and Area under PRC (AUPRC) was 0.97 and 0.951 (Fig. 3a, b). Taking nevi as a positive sample of the test set, the values of (AUROC, AUPRC) are (0.97, 0.951).

Then, as shown in Fig. 3c, the internal representations of melanoma and nevus were visualized by using tSNE method. The two diseases (melanoma is green, nevus is red) in the test set are separated perfectly by our model. This shows that CNN has learned the key features of the two classes.



**Fig. 3** CNN achieves high classification accuracy and can distinguish melanoma and nevus. **a** Receiver operating characteristic (ROC) curves of melanoma and nevus. The red line represents melanoma classification result, and the blue one represents nevus classification result. The star marks the best trade-off point. **b** Precision-recall curves (PRC) after classification of melanoma versus nevus in mixed magnification. **c** The tSNE image of melanoma (green point) and nevus (red point). After the features extraction of the trained CNN, melanoma and nevus naturally cluster in different clusters

## 4 Discussion

This study is the first deep learning based histopathological melanoma classification algorithm for Chinese patients. We collected a large Chinese dataset of 633 WSIs of melanoma and three nevi. And a novel CNN model was built and trained to identify melanoma with an AUROC of 0.97. As the [16, 17] shown, there is 25% discordance between pathologists in the histopathological diagnosis of melanoma [9–11]. The performance of model (accuracy of 92.5%, discordance of 7.5%) is on par with the performance of pathologists.

Deep learning may serve as the pathologist's eyes in the future. In this work, CNN was able to make diagnosis from a single patch, which would be considered as containing insufficient diagnosis information by pathologists. This comparable

performance of CNN may be explained by the ability of CNN to extract histopathological features from WSIs that are not detected by pathologists. Because the way CNN recognized images is different from that of humans. These visual image features provide opportunities for better quantitative modeling of disease, and it is possible to provide an efficient diagnosis method. In addition, this is also an opportunity for pathologists to gain insight into the features of melanoma deeply.

There are still some limitations to our current work that should be explored in future works. Studies have shown that additional clinical data can slightly increase the specificity and sensitivity of physician diagnosis. If other clinical data outside of pathological WSIs can be obtained during the clinical diagnostic process. Those additional clinical data may also be useful for model prediction, in the deep learning approach.

## 5 Conclusion

For the first time, deep learning was applied to melanoma classification of Chinese patients. We built a melanoma histopathology dataset with Chinese patients, and proposed and trained a novel CNN model to achieve high performance in the melanoma identification task. The model can achieve a diagnostic efficacy comparable to that of a pathologist. Conclusively, CNNs indicate to be a valuable tool to assist pathologists in diagnosing melanoma.

## References

1. Christopher. P.W., Bernard, W.S.: World Cancer Report 2014. World Health Organization, Geneva, Switzerland (2014)
2. Schadendorf, D., van Akkoi, J., Berking, C.: Melanoma. *The Lancet* **392**(10151), 971–984 (2018)
3. Intraocular, B.: Melanoma Treatment (PDQ): Health Professional Version. PDQ Cancer Information Summaries (2015)
4. Kurland, B.F., Gerstner, E.R., Mountz, J.M.: Promise and pitfalls of quantitative imaging in oncology clinical trials. *Magn. Reson. Imaging* **30**(9), 1301–1312 (2012)
5. Waldman, A.D., Jackson, A., Price, S.J.: Quantitative imaging biomarkers in neuro-oncology. *Nat. Rev. Clin. Oncol.* **6**(8), 445 (2009)
6. O'Connor, J.P.B., Jackson, A., Asselin, M.C.: Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. *Lancet Oncol.* **9**(8), 766–776 (2008)
7. Spratlin, J.L., Serkova, N.J., Eckhardt, S.G.: Clinical applications of metabolomics in oncology: a review. *Clin. Cancer Res.* **15**(2), 431–440 (2009)
8. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* **1**(5), 236–245 (2019)
9. Litjens, G., Sánchez, C.I., Timofeeva, N.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**(1), 1–11 (2016)

10. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyo, D.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**(10), 1559–1567 (2018)
11. Gecer, B., Akso, S., Mercan, E., Shapiro, L.G., Weaver, D.L.: Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recogn.* **84**, 345–356 (2018)
12. Hekler, A., Utikal, J., Enk, A.H., Berking, C., Klode, J.: Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur. J. Cancer* **115**, 79–83 (2019)
13. Wang, L., Ding, L., Liu, Z.: Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *Br. J. Ophthalmol.* **104**(3), 318–323 (2020)
14. Kulkarni, P.M., Robinson, E.J., Pradhan, J.S.: Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. *Clin. Cancer Res.* **26**(5), 1126–1134 (2020)
15. Association, GAoWM.: World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *The J. Am. Coll. Dent.* **81**(3), pp. 14–18 (2014)
16. Lodha, S., Saggar, S., Celebi, J.T., Silvers, D.N.: Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. *J. Cutan. Pathol.* **35**(4), pp. 349–352 (2008)
17. Corona, R., Mele, A., Amini, M., De Rosa, G.: Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. *J. Clin. Oncol.* **14**(4), 1218–1223 (1996)

# Handwritten Digits Recognition Based on Water Drop Algorithm and CNN



Geying Liang , Han Long , and Baoliang Dong

**Abstract** Handwritten digits recognition is an important research in pattern recognition and artificial intelligence, and it has broad application prospects in data processing. With the continuous development of deep learning, handwritten digits recognition technology has been widely improved. This work investigates the potential of the water drop segmentation algorithm and CNN in handwritten digits recognition, via deliberating the experiments on MNIST dataset. After selecting the appropriate starting point, specifying the moving rules and determining the direction, the water drop segmentation achieves excellent performance, especially for adhesion characters. The self-built neural network based on VGG NET which is realized with PyTorch framework is more targeted at the low-level features of the image and avoids the influence of the high-level feature factors on the pattern recognition results. Changing the multiple parameters of the network, this research is explicitly manifested, achieving the recognition accuracy of 0.99 on the handwritten digits data set MNIST.

**Keywords** Digits recognition · Water drop algorithm · CNN

## 1 Introduction

It is significant in the era of data that people write, use, transmit and filter out useful information as there is an explosive number of messages to extract and exchange. Such a collection, segmentation and recognition of digits may be displayed in a

---

G. Liang · B. Dong

China Electronics Technology Corporation 15, Beijing 100000, China

G. Liang

College of Electronic Science, National University of Defense Technology, Changsha 410000, China

e-mail: [lianggeying@126.com](mailto:lianggeying@126.com)

H. Long

College of Liberal Arts and Sciences, National University of Defense Technology, Changsha 410000, China

graphical way. Handwritten digits recognition is a detailed classification of Optical Character recognition [1]. As an important research in the field of pattern recognition and artificial intelligence, it is of great significance both on paper and in reality. With the rapid development of deep learning, handwritten digits recognition can achieve higher accuracy with the help of neural network. For the image recognition with noise and adhesion, traditional segmentation methods, such as projection method and CFS, are difficult to segment, which will affect the subsequent recognition. With simulating the process of water drop dripping, water drop segmentation algorithm is often used in character segmentation of images to cut the outline of the scene and is suitable for cutting the outline.

In order to solve the problem of handwritten digits recognition, *Yann LeCun* [2] proposed convolutional neural network named *LeNet* in 1998. For a long time, CNN maintained the best results on small scale issues like handwritten numbers in the world. *AlexNet* [3] was proposed by *Hinton* and *Alex* which improves the recognition accuracy to a higher level and arouses extensive attention of researchers to deep learning. It is the first successful application of tricks, such as ReLU, dropout and LRN and achieves the best classification in ILSVRC. Based on *AlexNet*, *VGG NET* [4] was proposed in 2014 that a smaller convolution kernel was used to deepen the network. ResNet was also put forward to reduce a series of problems brought by deep network, such as gradient disappearance. However, with the deepening of layers, the excessive consumption of computing resource increases the difficulty of training and causes non-convergence and gradient disappearance. To address this limitation, we propose to reconstruct the neural network based on VGG NET in image recognition. This research intends to experimental analyses of the water drop algorithm and CNN for effective improvement in handwritten digits recognition.

## 2 Methodology

### 2.1 *Character Segmentation*

Before image recognition and prediction, character segmentation should be carried out. The commonly used character segmentation methods include average algorithm, vertical-projection-based algorithm, connected-component-based algorithm and water drop algorithm.

**Vertical-projection-based algorithm.** The premise of the projection segmentation algorithm is that the image needs to go through the gray binarization. After binarization, there are only points with logic value of 0 and 1 in the image. The vertical projection method is used to statistically classify pixels in the same column, because the logical value of 1 between word intervals is far less than the value of 0 so that we can distinguish the peak and valley values and obtain the boundary.

**Table 1** Drop point adjacent pixels

N1	N0	N4
N2	N3	N5

This method is simple and effective, and is suitable for uncomplicated text verification codes, but has poor segmentation ability for images with noise or adhesion characters.

**Connected-component-based algorithm.** To visually show the connected domain segmentation method, connected-component-based algorithms often adopt the form of color filling segmentation algorithm (CFS). The split object of CFS is also a binary image. Assuming that all characters in the image are individually connected domains, find the non-zero point in the image as the starting point. After traversing the whole connected domain, we record the traversal points, and select a color to fill. If the traversal is completed, the next untraversed non-zero is found as the starting point, and then repeat traversal and filling until there is no untraversed non-zero in the whole image. CFS can clearly display the segmentation effect with different colors, but it cannot handle adhesion characters.

**Water drop segmentation algorithm.** Water drop segmentation algorithm splits adhesion characters by simulating the process of water drop dripping. The factors affecting water drop algorithm performance include starting point, moving rule and direction. Water drop up or down from the left and right sides of the string, depending on the starting point movement rules and the determined direction. Therefore, the water drop algorithm includes four segmentation methods.

The water drop segmentation algorithm needs to set several rules, which can be expressed by a matrix. For example, set two-dimensional array representing the matrix A and A [0] [5] (N0) is drop point of the algorithm. With the values of A[0][0] (N1) through A[5][6](N5) equal x, the dividing line is A [0] [5] (N0) to A [5] [5](N3). As shown in the table below. Water drop segmentation algorithm is often used in character segmentation of images to cut the outline of the scene. This method is suitable for cutting the outline, while it is subtle to choose the cut point (Table 1).

## 2.2 Convolutional Neural Networks

**Artificial Neural Network.** Artificial Neural Network (ANN) is a concept in Artificial intelligence. It uses machines to imitate the neural structure in the human brain and abstract into a model to deal with problems. According to the structure of human brain, artificial neural network composed of several neurons has a certain learning ability. It can evaluate the various parts of a function in parallel, without describing

the specific tasks of each unit. Artificial neural networks are the basis of most deep learning models. The neurons in the artificial neural network are divided into four parts: node, input, output, weight and bias. In the process of increasing processing experience, deep learning modifies the established mathematical model by correcting the weight of neurons.

**Convolutional Neural Network.** Convolutional Neural Network (CNN) as a feedforward artificial Neural Network is suitable for image processing whose neurons can extract features by rolling in different regions. The study of convolutional neural networks began in 1962, when *David Hubel and Torstein Wiesel* at Harvard University's Biological Neuroscience Laboratory proposed to record bioelectrical changes in a single neuron of cat's brain. They systematically described the structure of the visual cortex by recording single neurons. The concept of convolutional neural network had not been put forward by this experiment, but their methods laid a foundation for later research. Fukushima proposed a neural network structure including convolutional layer and pooling layer in the 1980s, and proposed the concept of Neurocognition. This became a milestone in the development of convolutional neural networks.

In general, the standard convolutional network consists of a convolutional layer, a nonlinear layer, a pooling layer and a lower sampling layer. Convolutional neural networks are especially suitable for neural networks in computer vision because they use local operations for hierarchical abstraction of representations. Instead of using one-to-one connections between all pixel units, convolutional Neural Network use grouped local connections as the first design idea. The second is the reliance on feature sharing, where each channel is generated by convolution using the same filter at all locations. It is two key design ideas that promote the success of convolutional architecture in the field of computer vision.

### 3 Experiment

#### 3.1 Datasets

We validate our network on MNIST. MNIST dataset was published by the National Institute of Standards and Technology. The training set consists of handwritten numbers from 250 randomly selected people, 50% of whom are high school students and the rest are the Census Bureau workers. The test set is also handwritten numeric data in the same proportion. The whole data set is composed of four parts with different functions but the same data type, namely training, training-target, test and test-target respectively. These four parts play different roles in training the neural network and testing the recognition accuracy of the network. In this study, the data set was divided into training set and test set according to the ratio of 6:1.

### 3.2 Image Processing

In this study, the difficulty of image processing lies in the character segmentation of handwritten digits, so we chose the water drop algorithm to implement it. Since the segmentation object of the character segmentation algorithm is a binary image, we use OpenCV to grayscale and binarization the image of the data set (Fig. 1).

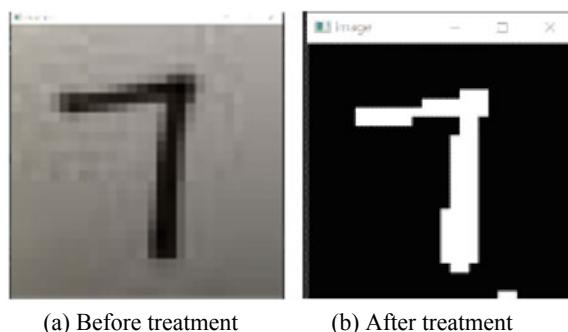
Then the binary image is segmented by characters. After the character histogram is obtained, the initial drop point of water droplets can be found, and the following rules were used to make a judgment:

- *Rule 1:* If there is only 0 or only 1 in the whole matrix representing only the handwritten part or the background part, the next drop point of the cut will be N3.
  - *Rule 2:* If the logic value of N2 is 1 for the existence of the handwriting with the other N1-N5 points having the only one which value equals 1, the next drop point of the cut will be N2.
  - *Rule 3:* If the logic value of N2 is 1 and N3 is 0 belonging background part, the next drop point of the cut will be N3.
  - *Rule 4:* If the logic value of both N2 and N3 equals 1 with N5 in background part, the next drop point of the cut will be N4.
  - *Rule 5:* If N2,N3 and N4 all belong to handwritten part, while N1 is the point of logic 0, the next drop point of the cut will be N5.
  - *Rule 6:* If in a pixel matrix, only N1 is logic 0, while all other points are pixels with logic 1, the next drop point of the cut will be N1.

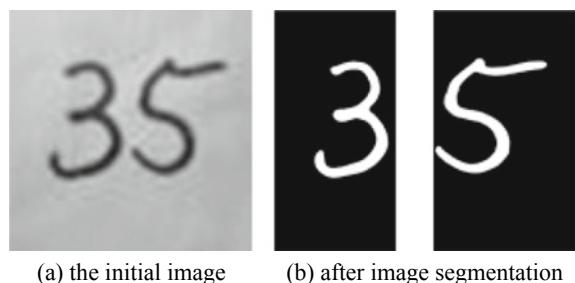
These rules essentially allow the water droplets to fall according to the natural rules that determine the curve of the cut. However, there is a deadlock in Rules 5 and 6, and we use the obstacle setting method to solve this problem. That is to detect the moving point, to observe if the moving point has been traversed. If it is not traversing, it will proceed normally; while it is traversing, we will put a barrier in it to prevent the droplet from moving.

The effect of the treatment is as follows. The two characters can be divided into two images respectively, which is ready for the next step of image recognition (Fig. 2).

**Fig. 1** Grayscale binarization effect



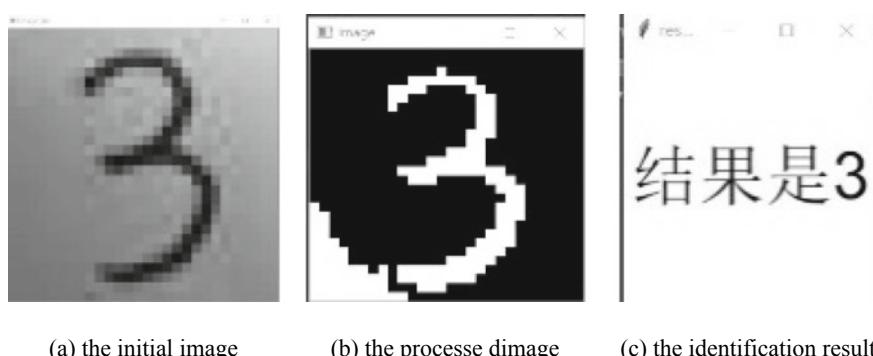
**Fig. 2** Character segmentation effect



### 3.3 Image Recognition

Handwritten digits recognition is based on the classification of image data into 1–9 Arabic numerals, which has a high fit with convolutional neural network. One use of CNN is to input an image and output results of classification. This work chooses to reconstruct the neural network based on VGG NET in image recognition. The self-built network structure can be more convenient to achieve the experimental goal.

The self-built network consists of input layer, convolution layer, pooling layer and full connection layer which are implemented by PyTorch framework. Compared with VGG NET, it has fewer convolution layers, which is more targeted at the low-level features of the image and avoids the influence of the high-level feature factors on the pattern recognition results. At the same time, the network structure adopts a 5\*5 convolution kernel improving the receptive field to a certain extent and obtaining more accurate global features than the former network. In addition, the network structure uses fewer layers to avoid the problem of large increase in computation caused by the enlargement of convolution kernel in ordinary multi-layer network structure. Moreover, the ReLU function is selected as the activation function, and the NLL\_LOSS function is selected as the loss function (Fig. 3).



**Fig. 3** Visual display of the recognition process

## 4 Result and Analysis

In the process of research, we design and carry out experiments on the best values of each parameters including batch-size, iteration times and learning rate, etc. First of all, 6 values are selected to determine the reasonable range of batch-size as the iteration is fixed at 4. The results are shown in the Table 2:

It can be seen that with the increase of batch-size value, the training time decreases continuously, while the precision increases first and then decreases. Considering the training time and identification accuracy, the batch-size value in study is set between 150 to 200.

Iteration times will directly affect the recognition accuracy of neural network in deep learning. Set epoch too small may lead to underfitting, on the contrary not only cause overfitting, but also increase the training time. In this study, when the batch-size is fixed at 128, the number of iterations is 1–18. The experimental results are shown in the Table 3:

According to several experiments, the optimal number of iterations is 16. At this time, 9892 digits pictures of the test set can be accurately identified out of the 10,000 digits pictures.

Learning rate can be understood as the descending speed in the process of gradient descent. If the learning step is too large, it may miss the lowest point of the loss function. However, the learning step set too small may cause the neural network stop at a local minimum rather than the expected global minimum. The influence of learning rate on identification accuracy is shown in the following Table 4:

**Table 2** Influence of different batch-sizes on accuracy and training time

batch-size	50	100	150	200	250	300	350
accuracy	95%	97%	98%	98%	97%	97%	96%
train-time	2min58s	2min43s	2min29s	2min20s	2min17s	2min13s	1min50s

**Table 3** Influence of different batch-sizes on accuracy and training time

iteration	1	2	3	4	5	6	7	8	9
accuracy	95%	97%	97%	98%	98%	99%	98%	99%	99%
iteration	10	11	12	13	14	15	16	17	18
accuracy	99%	98%	99%	99%	99%	99%	99%	98%	98%

**Table 4** Influence of learning rate on identification accuracy

learning rate	0.005	0.01	0.02	0.1	0.3	0.5
accuracy	96%	99%	98%	98%	97%	10%

To sum up, when batch-size is 128, the number of iterations is 16 and the value 0.01 is selected as the learning rate, the recognition accuracy reaches the best that 99% of handwritten digits images in the data set are successfully recognized.

## 5 Conclusion

This work has studied the prospective inference of the water drop segmentation algorithm and CNN in handwritten digits recognition, via deliberating the experiments on MNIST dataset. After the appropriate starting point is selected and the moving rules and determining the direction is specified, the water drop segmentation has achieved an excellent performance through the experiment. The self-built neural network based on VGG NET which is realized with PyTorch framework is more targeted at the low-level features of the image and avoids the influence of the high-level feature factors on the pattern recognition results. At the same time, the convolution kernel setting of 5\*5 is adopted, which can improve the receptive field to a certain extent and obtain more accurate global features. By varying the multiple parameters of the network, these experiments show that the optimum accuracy is 0.99.

## References

1. Nanehkaran, Y.A., Zhang, D., Salimi, S., et al.: Analysis and comparison of machine learning classifiers and deep neural networks techniques for recognition of Farsi handwritten digits[J]. *J. J. Supercomput.* **77**(2) (2021)
2. Lecun, Y., Bottou, L.: Gradient-based learning applied to document recognition[J]. *Proc. IEEE* **86**(11), 2278–2324 (1998)
3. Hinton, G.E., Srivastava, N., Krizhevsky, A., et al.: Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science* **3**(4), 212–223 (2012)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition[J]. *Computer Science* (2014)
5. Cecotti, H., Belad, A.: Rejection strategy for Convolutional Neural Network by adaptive topology applied to handwritten digits recognition. In: Eighth International Conference on Document Analysis & Recognition. IEEE Computer Society (2005)
6. Cireşan, D.C., Meier, U., et al.: Deep, big, simple neural nets for handwritten digit recognition.[J]. *Neural Comput. ation* (2010)

# A New Approach Based on Crater Detection and Matching for Self-Localization During Lunar Landings



Zhouyuan Qian , Hao Cheng , Tao Hu , Tao Cao , Yu Han , and Liang He

**Abstract** This paper provides an approach of visual localization based on crater detection and matching in lunar landing missions. Firstly, considering that the lunar image intensity is not uniform, and the contrast between the bright and dark areas is not obvious, a multi-threshold segmentation method based on gray features and geometric constraints is introduced to realize automatic extraction of multi-scale craters. Secondly, by constructing a rigid transformation model, the craters in descent images are roughly matched with the corresponding craters in database. Meanwhile, an efficient method using re-projection errors is proposed for reducing false matches so as to improve the localization accuracy. Images taken by Chang'e-3 landing camera and Lunar Reconnaissance Orbiter (LRO) are utilized to test the performance of the proposed approach, simulation results show that the proposed approach is able to detect the small-scale craters in low contrast areas with a high detection rate, and it can also match the correct craters to the database even when the resolution ratio between descent image and database is up to 10:1, which confirms that the proposed approach has strong robustness and high reliability.

**Keywords** Crater extraction · Multi-threshold segmentation · Image matching · Visual localization

## 1 Introduction

To realize high-precision lunar landing is an important task in manned lunar exploration missions, and since the inertial measurement unit (IMU) often suffers from time

---

Z. Qian · H. Cheng · T. Hu · T. Cao · Y. Han · L. He  
Shanghai Aerospace Control Technology Institute, Shanghai 201109, China  
e-mail: [zoe-qian@sjtu.edu.cn](mailto:zoe-qian@sjtu.edu.cn)

Shanghai Key Laboratory of Aerospace Intelligent Control Technology, Shanghai 201109, China

drifting, vision-based localization systems are quite popular nowadays. Considering that craters are widely distributed on the lunar surface and largely constant in shape, it's a promising approach to solve spacecraft's positioning problem by identifying impact craters [1].

To realize visual localization based on crater recognition, crater needs to be extracted from the images taken during the descent phase and matched with the same crater in the pre-loaded database [2]. At present, the commonly used crater extraction methods can be roughly divided into three types: terrain analysis method [3, 4], machine learning method [5, 6] and morphological fitting method [7, 8]. Zhou et al. [9] proposed a method to obtain true crater boundaries by extracting higher change rate of slope of aspect values at crater rims, yet this method was easy to miss detections of small-diameter shallow craters. DeLatte et al. [10] introduced a new CNN named Crater U-Net for the segmentation component of Mars craters, yet the detection result is sensitive to training data. As for crater matching problem, crater shape information, shadow information, template matching, or geometrical configurations are often used [11]. Woosang et al. [12] proposed a crater triangle matching algorithm, using invariants descriptors to match crater triangles. Hannah et al. [13] proposed a binary shadow matching approach by extracting and describing shadow features from lunar images, yet this approach had a limitation to provide the initial pose estimation of the spacecraft.

Considering the fact that intensity distribution varies strongly over the whole image, which increases the difficulty for crater segmentation and detection, we present a new approach of crater detection and matching algorithm mainly based on gray features and geometric constraints. Section 2 introduces the crater detection algorithm. In Sect. 3, the crater matching algorithm is presented. In Sect. 4, performance of proposed approach is validated through real lunar images from Cheng'e-3 and LRO.

## 2 Crater Detection Algorithm

### 2.1 Multi-Threshold Segmentation

Since lunar image intensity is not uniform, using a fixed threshold for binary segmentation might result in quantities of dark/bright areas lost in background areas. Hence, we use a dynamic changing threshold to realize the aggregation and growth of the bright and dark areas in the iteration process.

Specifically, the running threshold  $t_d$  changes iteratively within  $[t_{\min}^d, t_{\max}^d]$ , and the dark area candidates consist of the regions with pixel gray values smaller than  $t_d$ . Ideally, the dark areas appear as well-shaped and fully filled crescents. According to this geometric feature and the characteristic that the gray value of the dark area is lower than that of the surrounding background area, the dark area candidates are screened with constraints of fitting degree, saturation and grayscale contrast.

- (1) Fitting degree: The outer contour of each candidate area is extracted and fitted into a circle, and the requirement of fitting degree is met when

$$\begin{cases} f_1/f_o \geq c_1 \\ f_1/f_L \geq c_2 \\ 1 - f_2/f_L \geq c_3 \end{cases} \quad (1)$$

where  $f_o$  is the contour length,  $f_L$  is the circumference of the fitting circle,  $f_1$  is the number of pixels in outer contour which are successfully fitted while  $f_2$  is the number of pixels in outer contour which failed to fit the circle,  $c_1$ ,  $c_2$ ,  $c_3$  are three positive constants. The reliability of fitting is characterized by  $f_1/f_o$  to prevent under-fitting, and  $f_1/f_L$ ,  $1 - f_2/f_L$  are used to measure the fitting effect to filter the thin strip candidate area.

- (2) Saturation: The region size of the area bounded by the outer contour is  $s_t$  while that of the candidate area is  $s_r$ , and that of the fitting circle is  $s_f$ . The requirement of saturation is met only when

$$\begin{cases} s_r/s_t \geq c_4 \\ s_r/s_f \geq c_5 \end{cases} \quad (2)$$

where  $c_4$ ,  $c_5$  are positive constants.

- (3) Grayscale contrast: The grayscale average of the candidate area is  $g_r$  while that of its neighborhood is  $g_n$ , and the requirement of grayscale contrast is met only when

$$g_r/g_n \leq c_6 \quad (3)$$

where  $c_6$  is a positive constant.

During the iterations, the threshold  $t_d$  is gradually increased, and the dark candidate area aggregates and grows accordingly. When  $t_d = t_{\max}^d$ , the iteration result is obtained as the final distribution of dark areas.

As for bright areas, the running threshold  $t_b$  changes iteratively from  $t_{\max}^b$  to  $t_{\min}^b$ , and the bright area candidates consist of the regions with pixel gray values larger than  $t_b$ . The requirements of fitting degree and saturation are the same while the requirement of grayscale contrast should be

$$g_b/g_n \geq c_7 \quad (4)$$

where  $g_b$  is the grayscale average of the bright area candidate.

## 2.2 Crater Extraction

Having obtained the distribution of bright and dark areas, the crater extraction task can be reduced to finding the best fit regarding a fitness function  $h$ , which can be constructed by four factors:

- (1) Angle factor  $h_a$ [2]: We set the center of paired shadow area and bright area as  $\mathbf{C}_d(x_d, y_d)$  and  $\mathbf{C}_b(x_b, y_b)$  respectively. By connecting the center of each paired area, a crater vector is generated by  $\vec{C} = \mathbf{C}_b - \mathbf{C}_d$ . Since the angular deviation of  $\vec{C}$  from the image illumination direction  $\vec{S}$  should be close to zero,  $h_a$  can be obtained by:

$$h_a = 1 - \frac{\theta}{180} \quad (5)$$

where  $\theta = \arccos\left(\frac{\vec{C} \cdot \vec{S}}{|\vec{C}| |\vec{S}|}\right)$ ,  $|\vec{S}| = 1$ .

- (2) Grayscale factor  $h_g$ : This factor is used to measure the grayscale contrast between the bright and dark areas:

$$h_g = \frac{g_d}{g_b} \quad (6)$$

- (3) Distance factor  $h_d$ : It is the Euler distance between  $\mathbf{C}_d$  and  $\mathbf{C}_b$ . In this way, the paired bright and dark areas are kept close together.
- (4) Geometric factor  $h_s$ : Each paired bright area and dark area are gathered together and fit into a circle, the region size of the fitting circle is  $S_c$ , the region size of the bright area is  $S_b$  and the region size of the dark area is  $S_d$ .  $h_s$  can be obtained by

$$h_s = \frac{\min(S_b, S_d)}{\max(S_b, S_d)} \cdot \frac{S_b + S_d}{S_c} \quad (7)$$

The fitness function  $h$  is constructed by using the multiplication model, and it is defined as:

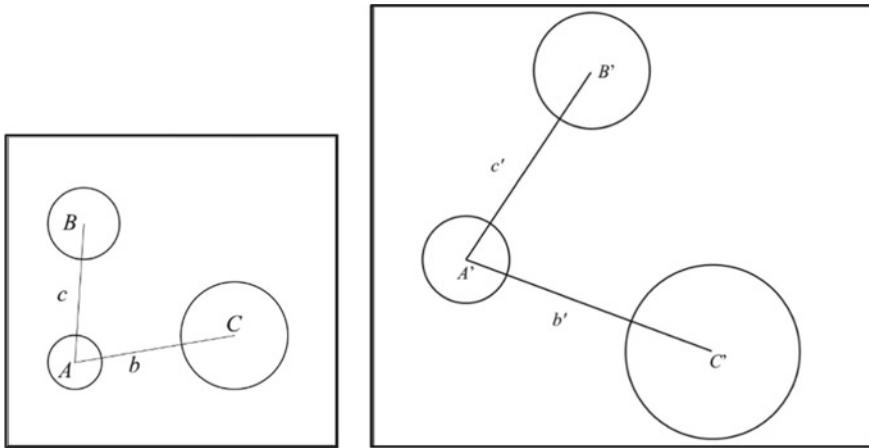
$$h = \frac{h_a \cdot h_s}{(h_d + \varepsilon) \cdot (h_g + \varepsilon)} \quad (8)$$

where  $\varepsilon$  is a positive constant. By finding the best fit regarding  $h$ , each paired bright area and dark area are found and correspondingly gathered together, and then the final edge of the crater can be obtained. The least square method is used to fit the edge into a circle to obtain the center and radius of the crater.

### 3 Crater Matching Algorithm

Images from Chang'e-3 landing camera can be obtained from <https://moon.bao.ac.cn>. The distortion of the images has already been corrected, and since the images taken in the vertical descent process are orthoimages, the transformation between Chang'e-3 descent images and LRO database can be reduced to the rigid transformation, and the scaling factor  $k$  can be approximated according to the image resolution. In this way, we can realize crater matching by using geometrical configurations. Figure 1 shows the rough crater matching diagram based on geometrical configurations, and  $A, B, C, A', B', C'$  are the centers of fitting circles of craters,  $r_A, r_B, r_C, r_{A'}, r_{B'}, r_{C'}$  are the radii, the distance between  $A, C$  is  $b$ , the distance between  $A, B$  is  $c$ , the distance between  $A', C'$  is  $b'$ , the distance between  $A', B'$  is  $c'$ .

For crater  $A$  in descent image, first any two neighboring centers  $B, C$  are selected, and under rigid transformation,  $\triangle ABC \sim \triangle A'B'C'$ , so ideally  $\angle A = \angle A'$ ,  $c' = k * c$ ,  $b' = k * b$ ,  $r_{A'} = k * r_A$ ,  $r_{B'} = k * r_B$ ,  $r_{C'} = k * r_C$ ,  $\frac{c'}{b'} = \frac{c}{b}$ . Considering the errors in crater detection process, the following formula can be established. For craters in descent image, find all the craters in database which satisfying the following formula, and all possible matching points are then constructed to realize the primary crater matching.



**Fig. 1** Crater matching diagram based on geometrical configurations

$$\left\{ \begin{array}{l} t_1 < \frac{\angle A'}{\angle A} < t_2 \\ t_1 < \frac{c/b'}{c/b} < t_2 \\ t_3 < \frac{c'}{k*c} < t_4 \\ t_3 < \frac{b'}{k*b} < t_4 \\ t_3 < \frac{r_{A'}}{k*r_A} < t_4 \\ t_3 < \frac{r_{B'}}{k*r_B} < t_4 \\ t_3 < \frac{r_{C'}}{k*r_C} < t_4 \end{array} \right. \quad (9)$$

where  $t_1, t_2, t_3, t_4$  are positive constants.

For each group of 3 pairs of matching points that satisfy formula (9), the affine transformation model as shown below is firstly used to calculate the transformation between the descent image and the database.

$$\begin{bmatrix} x^m \\ y^m \end{bmatrix} = \begin{bmatrix} a_0 & b_0 \\ a_1 & b_1 \end{bmatrix} \begin{bmatrix} x^r \\ y^r \end{bmatrix} + \begin{bmatrix} d_0 \\ d_1 \end{bmatrix} \quad (10)$$

$$\begin{bmatrix} x^r \\ y^r \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \\ a_3 & b_3 \end{bmatrix} \begin{bmatrix} x^m \\ y^m \end{bmatrix} + \begin{bmatrix} d_2 \\ d_3 \end{bmatrix} \quad (11)$$

where  $a_0, b_0, d_0, a_1, b_1, d_1, a_2, b_2, d_2, a_3, b_3, d_3$  are affine coefficients,  $x^r, y^r, x^m, y^m$  are the matching points in descent image and database.

Reprojection of all craters is carried out based on affine transformation model. Assuming that there's a crater  $i$  in descent image, the center of its fitting circle is  $\mathbf{X}_i^r$ , and its radius is  $r_i^r$ , and the center point is transformed to  $\mathbf{X}_i^{rr}$  in the coordinate system of database after reprojection. As for the crater  $j$  detected in database, the center of its fitting circle is  $\mathbf{X}_j^m$ , and its radius is  $r_j^m$ , and the center point is transformed to  $\mathbf{X}_j^{mr}$  in the coordinate system of the descent image after reprojection. When the formula below is satisfied, it is considered that the crater  $i$  and crater  $j$  meet the reprojection accuracy.

$$\left\{ \begin{array}{l} d(\mathbf{X}_i^{rr}, \mathbf{X}_j^m) \leq 1.5 * r_j^m \\ d(\mathbf{X}_j^{mr}, \mathbf{X}_i^r) \leq 1.5 * r_i^r \end{array} \right. \quad (12)$$

where  $d(\mathbf{X}, \mathbf{Y})$  is the function solving the Euler distance between  $\mathbf{X}, \mathbf{Y}$ . Assuming that the number of craters detected in descent image is  $n_r$ , the number of craters detected in database is  $n_m$ , and there're  $n$  craters that meet the reprojection accuracy. Only when  $n_r > 0.5 * n$ , the affine transformation model currently constructed is considered to be credible, and the crater pairs satisfying formula (12) are considered to be credible matching pairs.

For all matching points that satisfy formula (9), using the reprojection error to eliminate faulty matches. By establishing a matching counting matrix  $\mathbf{C}$  with a size

of  $(n_r \times n_m)$ ,  $\mathbf{C}(i, j)$  represents the number that crater  $I$  in descent image is matched with the crater  $j$  in database. When the formula below is satisfied, it is considered that crater  $i$  and crater  $j$  are the same crater.

$$\begin{cases} \mathbf{C}(i, j) = \max(\mathbf{C}(:, j)) \\ \mathbf{C}(i, j) = \max(\mathbf{C}(i, :)) \\ \mathbf{C}(i, j) > 0.7 * C_{\max} \end{cases} \quad (13)$$

where  $\max(\bullet)$  is a function solving the maximum value in the vector,  $C_{\max}$  represents the maximum value in matrix  $\mathbf{C}$ .

The matching counting matrix is used to extract all the matching point pairs of the two images, and the least square method is used to fit the optimal solution of affine transformation (10), so that realizing self-localization of the spacecraft in database.

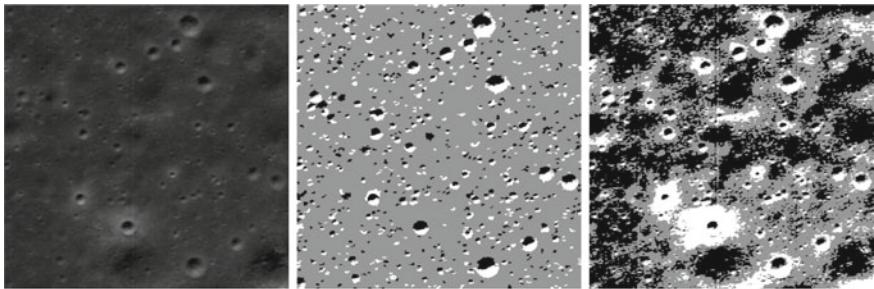
## 4 Experiments

### 4.1 Multi-Threshold Segmentation Experiment

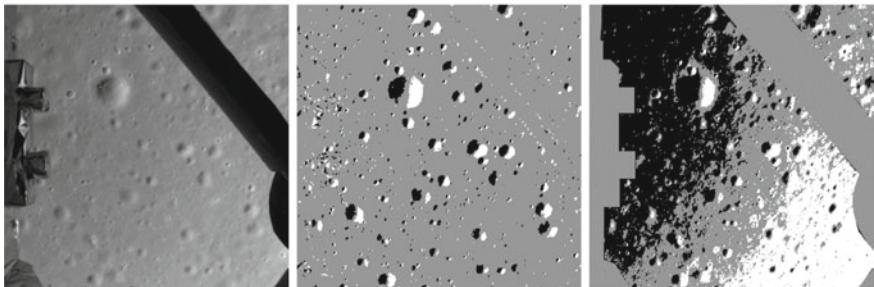
The validity of multi-threshold segmentation algorithm is proved on the LRO image and Chang'e-3 image, as shown in Fig. 2. Meanwhile, the crater extraction method based on the adaptive double-threshold segmentation algorithm adopted by Ref. [7] is selected as a comparison. As can be seen from the results, our algorithm can effectively solve the external influences such as illumination and weathering, while the adaptive double-threshold segmentation algorithm is unable to properly peel the bright and dark areas from the background area completely.

### 4.2 Crater Extraction Experiment

Two examples of the proposed crater detection algorithm are shown in Fig. 3. The detection rates of craters with a radius of larger than 4 pixels on LRO image and Chang'e-3 image are 96.7% and 87.6% respectively, showing strong robustness and high reliability. Meanwhile, the crater extraction results based on the adaptive double-threshold segmentation algorithm adopted by [7] are shown in Fig. 4, and the detection rates of craters with a radius of larger than 4 pixels on LRO image and Chang'e-3 image are 43.3% and 24.5% respectively, which confirms that the fixed threshold value can not adapt to the uneven intensity distribution of the lunar surface images.

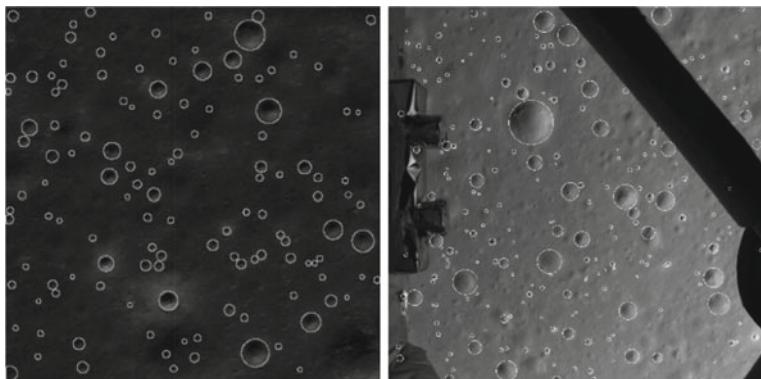


(a) LRO image, segmentation result of our algorithm, segmentation result of [7]



(b) Chang'e-3 image, segmentation result of our algorithm, segmentation result of [7]

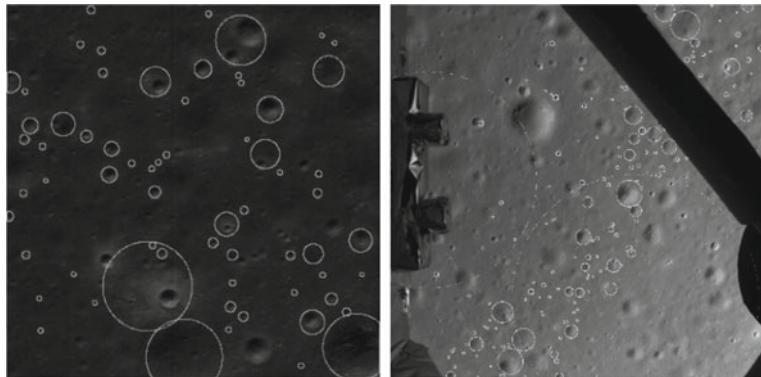
**Fig. 2** Segmentation results



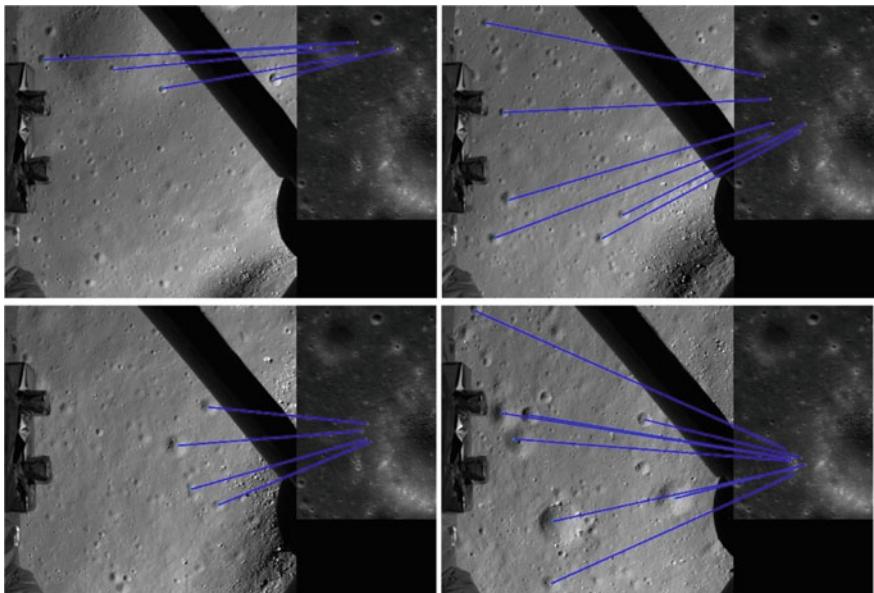
**Fig. 3** Crater detection results of our algorithm on LRO image and Chang'e-3 image

#### 4.3 Crater Matching Experiment

Using images from Chang'e-3 landing camera as the descent images and images from LRO as the database, experiments of absolute positioning of Chang'e-3 is performed based on our crater matching approach, and Fig. 5 shows four examples of the crater



**Fig. 4** Crater detection results of [7] on LRO image and Chang'e-3 image



**Fig. 5** Crater matching results

matching results. In each example, the left side shows Chang'e-3 image while the right side shows LRO database. The corresponding crater centers in these two images are connected by blue lines. In this experiment, the ratio of database resolution to descent image resolution decreased from 1.57 to 0.1, yet the correct matching of craters in these two images could still be realized. The average positioning error based on affine transformation model is about 1.53 pixels, which confirms that our proposed approach has wide applicability and high robustness.

## 5 Conclusion

In this paper, a new approach based on crater detection and matching for self-localization during lunar landings is proposed, the performance of proposed approach is tested through real lunar images from Chang'e-3 and LRO. Simulation results show that the automatic crater detection algorithm can effectively alleviate external influences such as illumination and weathering, and can comprehensively extract the small-scale craters with strong reliability. Furthermore, the proposed crater matching algorithm can achieve the correct matching even when the resolution ratio between the descent image and database is large, which confirms that our approach can still realize self-localization in an extreme case. In the future, the speed of the proposed approach needs to be further improved so as to realize real-time localization.

## References

1. Maass, B., Kruger, H., Theil, S.: An edge-free, scale-, pose-and illumination-invariant approach to crater detection for spacecraft navigation. In: 7th International Symposium on Image and Signal Processing and Analysis, pp. 603–608. IEEE, Dubrovnik, Croatia (2011)
2. Meng, Y., Hutao, C., Yang, T.: A new approach based on crater detection and matching for visual navigation in planetary landing. *Adv. Space Res.* **53**(12), 1810–1821 (2014)
3. Min, C., Danyang, L., Kejian, Q., et al.: Lunar Crater detection based on terrain analysis and mathematical morphology methods using digital elevation models. *IEEE Trans. Geosci. Remote Sens.* **56**(7), 3681–3692 (2018)
4. Bo, L., Zongcheng, L., Jiang, Z., et al.: Automatic detection and boundary extraction of lunar craters based on LOLA DEM data. *Earth Moon Planet.* **115**, 59–69 (2015)
5. Yanmin, J., Fan, H., Shijie, L., et al.: Small scale crater detection based on deep learning with multi-temporal samples of high-resolution images. In: 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images, MultiTemp 2019. IEEE, Shanghai (2019)
6. Delatte, D.M., Crites, S.T., Guttenberg, N., et al.: Automated crater detection algorithms from a machine learning perspective in the convolutional neural network era. *Adv. Space Res.* **64**(8), 1615–1628 (2019)
7. Wang, D., Xing, S., Xu, Q., et al.: A planetary image based automatic impact crater extraction method. *J. Astronaut. ICS* **36**(10), 1163–1171 (2015)
8. Pedrosa, M.M., De Azevedo, S.C., Da Silva, E.A., et al.: Improved automatic impact crater detection on Mars based on morphological image processing and template matching. *Geomat. Nat. Haz. Risk* **8**(2), 1306–1319 (2017)
9. Yi, Z., Hao, Z., Min, C., et al.: Automatic detection of lunar craters based on DEM data with the terrain analysis method. *Planet. Space Sci.* **160**, 1–11 (2018)
10. DeLatte, D.M., Crites, S.T., Guttenberg, N., et al.: Segmentation convolutional neural networks for automatic crater detection on mars. *IEEE J. Selected Topics Appl. Earth Observ. Remote Sens.* **12**(8), 2944–2957 (2019)
11. Shao, W., Xie, J., Cao, L., et al.: Crater matching algorithm based on feature descriptor. *Adv. Space Res.* **65**(1), 616–629 (2020)
12. Woosang, P., Youeyun, J., Hyochoong, B., et al.: Robust crater triangle matching algorithm for planetary landing navigation. *J. Guid. Control. Dyn.* **42**(2), 402–410 (2019)

13. Kaufmann, H., Lingenauber, M., Bodenmueller, T., et al.: Shadow-based matching for precise and robust absolute self-localization during lunar landings. In: 2015 IEEE Aerospace Conference, AERO 2015, pp. 1–13. IEEE, Big Sky, MT, United states (2015)

# **Intelligent Systems**

# Robust Spectral Clustering via the Ordering Metric



Bingjie Li , Tianhao Ni , and Zhenyue Zhang

**Abstract** Spectral clustering is one of the most popular algorithms in unsupervised learning. However, it is difficult to construct an affinity graph that benefits spectral clustering, mainly due to the lack of a discriminative distance metric. In this article, motivated by the weakness of the traditional distance metric, we propose a novel metric named the ordering metric, in order to measure the class-consistency of two data points. Based on the proposed metric, a scalable Gaussian affinity graph is constructed. The ordering metric can distinguish classes more accurately than classical metrics, hence our proposed affinity graph based on it can simultaneously highlight intra-class connections and suppress inter-class connections. With these advantages, the spectral approach normalized cut can achieve a low-dimensional projection from the graph that contains clear and correct class information. Classical clustering approaches such as K-means can cluster these projection points perfectly, due to the significant separation of the spectral projection between classes. Numerical experiments on 2 synthetic 2D data sets and 5 real-world data sets show the outstanding clustering performance of our algorithm.

**Keywords** Unsupervised learning · Spectral clustering · Affinity graph · Metric learning

## 1 Introduction

Clustering is a fundamental technique to retrieve the underlying class information of data. It has been widely adopted in a variant of applications such as face recognition [1], image segmentation [2], text analysis [3], medical diagnosis [4], and so on. Basically, Clustering aims to partition data into several groups based on the pairwise similarities of data points.

---

B. Li · T. Ni  
Zhejiang University, Hangzhou 310027, China

Z. Zhang ()  
Zhejiang Lab, Hangzhou 311122, CHINA and Zhejiang University, Hangzhou 310027, China  
e-mail: [zyzhang@zju.edu.cn](mailto:zyzhang@zju.edu.cn)

Some direct algorithms have been developed for data clustering in the early years. The well-known K-means [5] aims to partition data points so as to minimize the within-cluster sum of squares. However, it converges to a local optimum generally, closely relied on data distribution and initial center seed selection. Other direct clustering methods contain agglomerative clustering [6], mean-shift [7], Gaussian mixture model [8], and so on. Dimension reduction techniques, including principal component analysis [9], non-negative matrix factorization [10], and manifold learning methods [11], [12], [13], are commonly used to find clustering-friendly representations of data so that direct clustering approaches could be applied.

As a combination strategy of direct algorithms and dimension reduction, spectral clustering has been proven effective in various applications. Spectral clustering, also known as graph clustering, aims to learn an affinity graph highlighting intra-class connections and suppressing inter-class connections. Generally, the affinity of  $x_i$  and  $x_j$  is determined by the Gaussian function  $\exp(-d^2(x_i, x_j))$ , where  $d(x_i, x_j)$  is a metric between  $x_i$  and  $x_j$ . However, for real-world data sets, the commonly used Euclidean metric performs poorly, mainly because it cannot effectively identify the class-consistency of points at the junction of two classes, especially for data with different density distributions.

In this paper, inspired by the weaknesses of traditional clustering algorithms, we provide a novel graph-based clustering method entitled the ordering spectral clustering (OSC) method. The OSC method can be divided into four steps. In the first step, we propose a new metric, named as the ordering metric, to identify data at the junction of two classes with different densities. In the second step, a scalable Gaussian graph is constructed based on the ordering metric. The third step is to obtain a low-dimensional projection suitable for clustering by the normalized cut [2]. Finally, we obtain the clustering results by implementing K-means [5] on the spectral projection.

For the paper, the main contributions are as follows: (1) we propose a new metric that can identify the class of marginal samples. The metric ignores the distance value and re-scales the distance to the centroid point, which helps to distinguish classes with different densities. (2) The scalable Gaussian graph constructed in our paper simultaneously highlights intra-class connections and suppresses inter-class connections, which guarantees that spectral approaches can obtain a low-dimensional projection that contains clear and correct class information. (3) Due to the above advantages for the metric and the graph, our approach has superior performance in clustering, compared with 5 state-of-the-art clustering methods. Numerical experiments show that our approach has a much lower clustering error rate than other algorithms on a variety of real-world data sets with different scales, dimensions, and clusters.

## 2 The Ordering Metric

Given a data set  $X = \{x_i\}$  of  $n$  points, at each point  $x_i$ , we can reorder  $\{x_j\}$  as  $x_{i1}, \dots, x_{in}$  according to their Euclidean distances to the centroid point  $x_i$ . Equivalently, it assigns an ordering index for each  $x_j$  corresponding to  $x_i$ . More precisely, we define the ordering function as

$$o(x_j; x_i) = |\{x \in X, x - x_{i2} < x_j - x_{i2}\}|. \quad (1)$$

where  $|\cdot|$  means the number of points in the set. Specially,  $o(x_i; x_i) = 0$ . Here,  $x_i$  can be taken as a parameter vector of the function  $o(x; x_i)$ .

There are some special propositions of  $o(x; x_i)$  different from the Euclidean metric  $x - x_{i2}$ , fixing  $x_i$ . At first, the function  $o(x; x_i)$  ignores the distance value, or more precisely, it re-scales the distance to the centroid point. That is, a closed pair  $x_j$  and  $x_k$  may have a large gap  $|o(x_j; x_i) - o(x_k; x_i)|$ , and a relatively far pair  $x_j$  and  $x_k$  may have a not larger gap. Secondly,  $o(x; x_i)$  depends on the local density of the centroid point  $x_i$  very much. For example, if the neighboring points of  $x_i$  are denser than the neighboring points of  $x_j$  and  $x$  is between  $x_i$  and  $x_j$ , then  $o(x; x_i)$  is larger than  $o(x; x_j)$  generally.

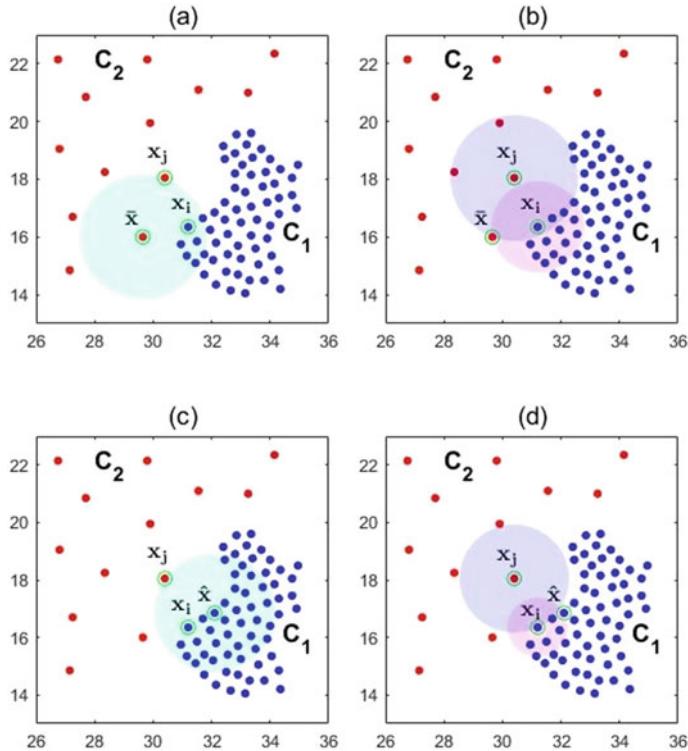
These propositions are helpful to distinguish classes with different densities. To show it, let us consider the simple example of two classes  $\mathcal{C}_1$  and  $\mathcal{C}_2$  in Fig. 1, where  $\mathcal{C}_1$  is denser than  $\mathcal{C}_2$ . Fixed  $x_i \in \mathcal{C}_1$  and  $x_j \in \mathcal{C}_2$ , we set  $\mathcal{N}(x_i)$  and  $\mathcal{N}(x_j)$  as two disk-neighbor sets of  $x_i$  and  $x_j$  with an overlap, respectively. We consider two point  $\bar{x}$  and  $\hat{x}$  in the intersection part, where  $\bar{x}$  is class-consistent with  $x_j$ ,  $\hat{x}$  is class-consistent with  $x_i$ .

By the ordering function, we can detect that  $\bar{x}$  is class-consistent with  $x_j$  rather than  $x_i$ , noticing that  $o(\bar{x}; x_j) < o(\bar{x}; x_i)$ , as in (b) of Fig. 2. Meanwhile, the class of  $\bar{x}$  can't be detected correctly by the Euclidean metric, since  $\bar{x} - x_{i2} < \bar{x} - x_{j2}$ , as in (a) of Fig. 2. However, identifying  $x$  by comparing  $o(x; x_j)$  and  $o(x; x_i)$  may be not precious to characterize  $x$ 's class for some  $x$ . To show this, let's consider  $x = \hat{x}$  in (c) and (d) of Fig. 2. If we detect the class of  $\hat{x}$  by comparing  $o(\hat{x}; x_j)$  and  $o(\hat{x}; x_i)$  as in (d) of Fig. 2, we will obtain a misleading result since  $o(\hat{x}; x_j) < o(\hat{x}; x_i)$ . However, if we consider  $x_i$  and  $x_j$  as two neighbors of  $\hat{x}$  as in (c) of Fig. 2, we will find that  $o(x_i; \hat{x})$  and  $o(x_j; \hat{x})$ , which matches the ground truth class of  $\hat{x}$ . In summary, we find that

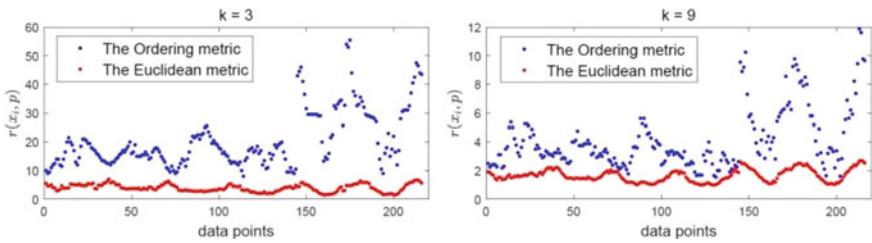
$$\phi(\bar{x}, x_j) < \phi(\bar{x}, x_i), \phi(\hat{x}, x_i) < \phi(\hat{x}, x_j). \quad (2)$$

where  $\phi(x_p, x_q)$  is defined as  $\max\{o(x_p; x_q), o(x_q; x_p)\}$ . Hence, the symmetric function  $\phi(x_p, x_q)$  can really reflect the relationship between  $x_p$  and  $x_q$ . We call  $\phi(\cdot, \cdot)$  as the **ordering metric**.

Given  $x_i$  and its  $k$ -th neighbor  $x_{i,k}^o$  under the ordering metric, we consider the  $k$ -nn neighborhood of  $x_i$  determined by the ordering metric  $\phi(x, x_i)$ :



**Fig. 1** **a:** Observing  $x_i$  and  $x_j$  as neighbors of  $\bar{x}$ . **b:** Observing  $\bar{x}$  as a neighbor of  $x_i$  and  $x_j$ , respectively. **c:** Observing  $x_i$  and  $x_j$  as neighbors of  $\hat{x}$ . **d:** Observing  $\hat{x}$  as a neighbor of  $x_i$  and  $x_j$ , respectively



**Fig. 2** The value of  $r(x_i, p)$  under the Euclidean metric and the ordering metric of each  $x_i$  in COIL-3 with  $k = 3$  (left) and  $k = 9$  (right)

$$\mathcal{N}_o(x_i, k) = \{x \in X : \phi(x, x_i) \leq \phi(x_{i,k}^o, x_i)\} \#(1)$$

It has two advantages as follows.

- The neighbors in each  $\mathcal{N}_o(x_i, k)$  are more likely closed to  $x_i$ , and also,  $x_i$  is a closer neighbor of each point in the set. It implies that  $\mathcal{N}_o(x_i, k)$  is more likely class-consistent with  $x_i$  even if the nearby classes have different densities.
- The boundary of the neighbor set  $\mathcal{N}_o(x_i, k)$  is distinguishable from  $x_i$ 's class-inconsistent set  $\mathcal{N}_C(x_i) = \{x \in X : \ell(x) \neq \ell(x_i)\}$ , where  $\ell(x)$  is the ground truth label of  $x$ .

To illustrate the above advantages, we provide a criterion to evaluate the quality of a given neighborhood  $\mathcal{N}_o(x_i, k)$ . Given  $x_i \in X$  and an integer  $k$ , we define

$$r(x_i, k) = \frac{\phi(x_i, \hat{x}_i)}{\phi(x_i, x_{i,k}^o)}, \#(2) \quad (3)$$

where  $\hat{x}_i$  is the closest class-inconsistent point of  $x_i$ .<sup>1</sup> In (2),  $\phi(x_i, \hat{x}_i)$  can be seen as the distance between  $x_i$  and  $\mathcal{N}_C(x_i)$ , and  $\phi(x_i, x_{i,k}^o)$  approximately represents the radius of  $\mathcal{N}_o(x_i, k)$ . Given a  $x_i \in X$  and an integer  $k$ , the larger  $r(x_i, k)$  is, the more distinguishable  $\mathcal{N}_o(x_i, k)$  is from  $\mathcal{N}_C(x_i)$ . Furthermore,  $\mathcal{N}_o(x_i, k)$  is class-consistent if and only if  $r(x_i, k) > 1$ .

Figure 2 shows the value of  $r(x_i, k)$  of each  $x_i$  in a real-world data set COIL-3<sup>2</sup> under the Euclidean metric and the ordering metric, with  $k = 3$  and  $k = 9$ . For each  $x_i$ , the ordering metric always provide larger  $r(x_i, k)$  than that of the Euclidean metric. Hence, the ordering metric separate  $\mathcal{N}_o(x_i, k)$  from other classes more distinguishable than the Euclidean metric, regardless of different  $k$ .

### 3 Graph Construction for Clustering

#### 3.1 A Scalable Affinity Based on the Ordering Metric

The construction of an affinity graph is a key step of spectral clustering. The affinity graph regards the data points as nodes, and the affinity between the points, that is, the probability of class-consistency, as the weight of edges. Generally, the low-dimensional spectral projection of the affinity graph maintains the local structure of the data and is more suitable for clustering than the original data.

The good separation provided by the ordering metric prompts us to construct a scalable affinity function:

$$A(x_j; x_i) = \exp\left(-\frac{\phi(x_j, x_i)^2}{\phi(x_{i,k}^o, x_i)^2}\right), \#(3) \quad (4)$$

---

<sup>1</sup> Similarly,  $r(x_i, p)$  can also be defined under Euclidean metric.

<sup>2</sup> COIL-3 is a subset of image data set COIL-20. The result of COIL-20 is reported in Section IV.

where  $k$  is a parameter to be set. The scaled affinity function  $\mathcal{A}(x_j; x_i)$  benefits clustering very much, due to the following two reasons.

- For each  $x_j$  in  $\mathcal{N}_o(x_i, k)$ ,  $\mathcal{A}(x_j; x_i) \geq 1/e$ . It implies that the connection between  $x_i$  and its class-consistent neighborhoods is strong enough.
- For each  $x_m$  that is class-inconsistent with  $x_i$ , we have

$$\mathcal{A}(x_m; x_i) \leq \exp\left(-\frac{\phi(\hat{x}_i, x_i)^2}{\phi(x_{i,k}^o, x_i)^2}\right) \leq \exp(-r(x_i, k)^2) \quad (5)$$

Hence, inter-class connections of  $x_i$  are effectively suppressed since  $r(x_i, p)$  is generally large.

### 3.2 Graph Construction and Spectral Clustering

Based on  $\mathcal{A}(\cdot; \cdot)$ , we construct a symmetric graph  $G$  with affinity as

$$g_{ij} = \min\{\mathcal{A}(x_i; x_j), \mathcal{A}(x_j; x_i)\} \#(4) \quad (6)$$

since the minimum of  $\mathcal{A}(x_j; x_i)$  and  $\mathcal{A}(x_i; x_j)$  weakens the inter-class connection. The affinity  $g_{ij}$  has similar properties with  $\mathcal{A}(x_j; x_i)$  and  $\mathcal{A}(x_i; x_j)$ . For one thing, it's easy to obtain  $g_{ij} \geq 1/e$  if  $x_j \in \mathcal{N}_o(x_i, k)$  and  $x_i \in \mathcal{N}_o(x_j, k)$ . For another, we have

$$g_{ij} \leq \min\{\exp(-r(x_i, k)^2), \exp(-r(x_j, k)^2)\} \quad (7)$$

if  $x_i$  and  $x_j$  are class-inconsistent. In our experiment, we observe that  $\exp(-r(x_i, k)^2)$  is extremely small for almost every  $x_i$  with a suitable parameter  $k$ . For instance,  $r(x_i, k)$  is larger than 8 for each  $x_i$  in COIL-3 when  $k = 3$ , which implies that each  $\exp(-r(x_i, k)^2)$  is smaller than  $10^{-27}$ . Hence, the proposed graph  $G$  is automatically sparse, if we ignore the extremely small entries of  $G$ .

The above observation implies that  $G$  simultaneously highlights intra-class connections and suppresses inter-class connections. This property helps spectral approaches to obtain a low dimensional projection that contains clear and accurate class information. The normalized cut (N-cut) [2] is a commonly used spectral projection method. Given an affinity graph  $G$ , It computes the eigenvectors corresponding to the smallest  $k$  eigenvalues of

$$L = I - D^{-1/2} G D^{-1/2} \quad (8)$$

where  $D$  is a diagonal matrix with  $d_{ii} = \sum_{j=1}^n g_{ij}$ . N-cut works very well for our graph. Figure 3 shows the first three eigenvectors of  $L$  of COIL-3. These eigenvectors clearly retrieve the class information of all data points. Classical approaches such as K-means can perfectly cluster these data points due to the significant separation between classes.

The computation of  $\phi(x_i, x_j)$  for all  $x_i$  and  $x_j$  cost totally  $O(n^2 \log n)$ . We simply reduce it to  $O(n^2 \log p)$  by setting  $\phi(x_i, x_j) = +\infty$  if  $o(x_i; x_j) > p$  or  $o(x_j; x_i) > p$ , where  $p$  is an integer parameter. In our experiment, we observe that such an approximation of  $G$  does not change the clustering result if we set  $p$  as 100. This is mainly because the value of  $g_{ij}$  is so small that can be ignored if  $\phi(x_i, x_j) > 100$ . Hence, we approximately compute the ordering metric of  $x_i$  and  $x_j$  b

$$\phi(x_i, x_j) = \begin{cases} \max\{o(x_i; x_j), o(x_j; x_i)\}, & o(x_i; x_j) \leq 100 \text{ and } o(x_j; x_i) \leq 100 \\ +\infty, & \text{otherwise.} \end{cases} \quad \#(5) \quad (9)$$

Notice that

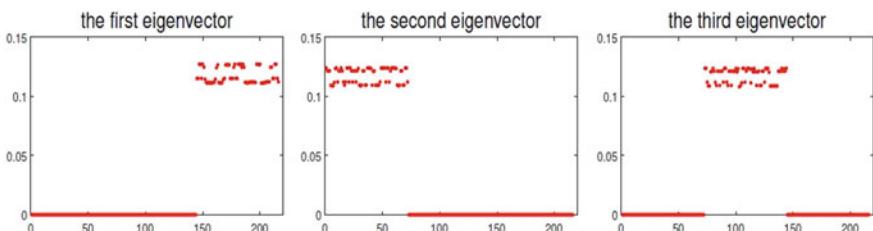
$$g_{ij} = \exp(-\infty) = 0 \text{if } \phi(x_i, x_j) = +\infty. \quad (10)$$

**Algorithm 1: Ordering Spectral Clustering (OSC)** **Input:**  $X = \{x_1, \dots, x_n\}$ ,  $K$ : the number of clusters,  $k \in \mathbb{N}$  the parameter in scaled affinity function

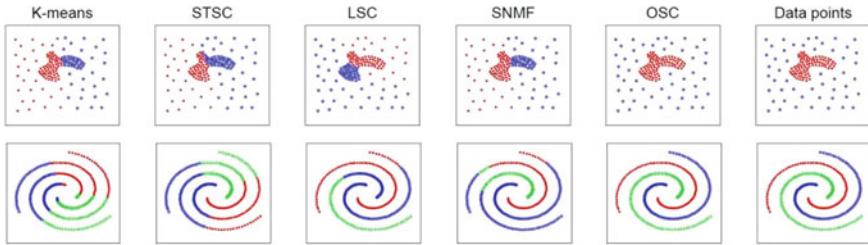
**Output:** A partition of  $X$ :  $X = \{X_1, \dots, X_K\}$

- 1: Calculate  $\phi(x_i; x_j)$  for each  $x_i, x_j \in X$  by (5)
- 2: Construct the affinity matrix  $G$  by (3) and (4)
- 3: Compute the first  $K$  eigenvectors  $u_1, \dots, u_K$  of  $L = I - D^{-1/2}GD^{-1/2}$ , and set  $U = [u_1, \dots, u_K]^T$
- 4: Use K-means to cluster the 2-norm normalized row vectors of  $U$  into  $K$  clusters:  $\{X_1, \dots, X_K\}$

The flow of our proposed ordering spectral clustering (OSC) is illustrated in Algorithm 1. The most time-consuming step in our algorithm is the eigenvalue decomposition of  $L$ , which costs  $O(n^3)$  per iteration. Fortunately, our graph has a sparse



**Fig. 3** The first 3 eigenvectors of  $L = I - D^{-1/2}GD^{-1/2}$  (COIL-3)



**Fig. 4** Comparison of the clustering results of K-means, STSC, LSC, SNMF and OSC (from left to right) on the two synthetic data sets Gourd and Spiral (last column, from top to bottom)

structure, which is suitable for fast eigenvalue decomposition algorithms, such as Krylov subspace method [14]. The experiments in the next section show that our algorithm has significant advantage in speed.

## 4 Numerical Experiments

In this section, we report the clustering performance of our proposed algorithm OSC, compared with five state-of-the-art algorithms for clustering, including K-means [5], STSC [15], LSC [16], EnSC [17] and SNMF [18]. A variety of data sets, including two 2D synthetic data sets and five real-world data sets were tested. All compared algorithms are executed on the Windows system in a PC with Intel Core i5-8250U CPU@1.60 GHz and 8 GB RAM.

### 4.1 Synthetic Data Sets

In this subsection, we test five clustering algorithms on two synthetic 2D data sets Compound and Spiral.<sup>3</sup> Each of them has special class patterns that may not benefit the detection of ground-truth classes. we show the performance of K-means, STSC, LSC, SNMF and our algorithm OSC on these data sets in Fig. 4.<sup>4</sup> The parameter  $k$  in OSC is set as 3. As in Fig. 4, none of the compared methods are able to achieve the correct clustering partition except OSC. OSC correctly recognized all the classes in the two sets, though the data distributions and densities are quite different.

<sup>3</sup> Compound and Spiral can be downloaded at <http://cs.joensuu.fi/sipu/datasets/>.

<sup>4</sup> The result of EnSC is omitted here since the subspace learning approach is not suitable for 2D data sets.

**Table 1** Clustering error rate (percentage) of the algorithms on real-world data sets

Data sets	K-means	STSC	LSC	EnSC	SNMF	OSC
COIL-20	35.69	49.72	19.79	<u>19.03</u>	26.94	<b>0</b>
COIL-100	52.17	55.17	44.60	<u>41.17</u>	44.75	<b>18.99</b>
ORL	32.00	<u>19.00</u>	19.75	22.50	23.50	<b>14.75</b>
Umist	58.05	58.05	<u>41.24</u>	55.57	49.56	<b>23.19</b>
Optdigit	20.81	23.70	<u>3.83</u>	15.30	12.37	<b>2.46</b>

**Table 2** Computational time (seconds) of the compared algorithms

Data sets	K-means	STSC	LSC	EnSC	SNMF	OSC
COIL-20	6.94	1.28	<u>0.70</u>	2.28	5.68	<b>0.47</b>
COIL-100	186.26	71.71	<u>14.11</u>	44.81	111.69	<b>11.34</b>
ORL	19.58	<u>0.64</u>	0.55	17.40	6.01	<b>0.52</b>
Umist	1.63	0.42	<u>0.28</u>	0.94	4.27	<b>0.20</b>
Optdigit	3.18	18.23	<b>0.86</b>	10.66	13.07	<u>3.01</u>

## 4.2 Real-World Data Sets

We conduct experiments on five real-world data sets, including COIL-20, COIL-100 [19], ORL [20], UMist [21] and Optdigit.<sup>5</sup> The five data sets have different scales, dimension, and clusters.

We list the clustering error rate in Table 1 for the OSC and other compared algorithms on the five data sets. In Table 1 and the following Table 2, the best results are boldfaced, and the second-best results are noted with underline. The parameter  $k$  of OSC in COIL-20, COIL-100, ORL, UMist and Optdigit is set as 3, 3, 4, 3, 10, respectively. The early baseline approaches such as K-means and STSC cannot give acceptable clustering results on these databases except that STSC performs well on ORL. SNMF performs slightly better than the early approaches, as it adopted sparse affinity graphs that suppressed inter-class connections. LSC and EnSC can also improve clustering accuracy, mainly because they effectively mine the subspace structure of the data sets by different self-expression strategies.

The OSC performs significantly better than all of other methods on the five data set. The lowest clustering error rate of the compared approaches on COIL-20, COIL-100, ORL, UMist and Optdigit can be decreased by 19.03%, 22.18%, 4.25%, 18.05% and 1.37%, respectively. The clustering result of COIL-20 is better than newly-developed deep network algorithms like DSC [22], DBC [23] and GALA [24]. We achieve a 0% error rate on COIL-20 while the error rate of COIL-20 reported in DSC, DBC and GALA are 5.14%, 20.70% and 20.00%, respectively.

Table 2 lists the computational costs of the OSC and other five algorithms. OSC costs least on four data sets and the second least on Optdigit. The result in Tables 1 and 2 shows that OSC has both better performance and high efficiency.

---

<sup>5</sup> Optdigit can be downloaded at <https://archive.ics.uci.edu/ml/datasets/>

## 5 Conclusion

In this paper, we proposed a novel metric that can distinguish classes more accurately than classical metrics. Based on our proposed metric, we develop a scalable graph construction algorithm for spectral clustering. Our graph could simultaneously highlight intra-class connections and suppress inter-class connections, resulting in superior performance in clustering. Further efforts are required for promoting our graph construction method to other machine learning tasks such as semi-supervised learning.

**Acknowledgements** The work was supported in part by NSFC project 11971430 and Major Scientific Research Project of Zhejiang Lab (No.2019KB0AB01).

## References

1. J. Wright, A.Y., Yang, A., Ganesh, S., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2), 210–227 (2009)
2. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* **22**(8): pp. 888–905 (2000)
3. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 267–273 (2003)
4. Soni, J., Ansari, U., Sharma, D., Soni, S.: Predictive data mining for medical diagnosis: An overview of heart disease prediction. *Int. J. Comput. Appl.* **17**(8), 43–48 (2011)
5. Lloyd, S.: Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **28**(2), 129–137 (1982)
6. Everitt, B.: Cluster analysis. *Quality Quantity: Int. J. Methodol.* **14**(1), 75–100 (1980)
7. Cheng, Y.: Mean shift, mode seeking, and clustering. *IEEE Trans. Patt. Anal. Mach. Intell.* **17**(8), 790–799 (1995)
8. Christopher, M.: Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag Berlin, Heidelberg (2006)
9. Jolliffe, I.: *Principal Component Analysis* (1986)
10. Daniel, D., Lee, H., Seung, S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791 (1999)
11. Sam, T.: Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
12. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for non-linear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
13. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.* **26**(1), 313–338 (2005)
14. Stewart, W. G. A.: krylov-schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.* (2001)
15. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. *Adv. Neural In-formation Process. Syst.* **17**, 1601–1608 (2004)
16. Cai, D., Chen, X.: Large scale spectral clustering via landmark-based sparse representation. *IEEE Trans. Cybern* **45**(8), 1669–1680 (2015)
17. You, C., Li, C.G., Robinson, D.P., Vidal, R.: Oracle based active set algorithm for scalable elastic net subspace clustering. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3928–3937 (2016)

18. Zhu, Z., Li, X., Liu, K., Li, Q.: Dropping symmetry for fast symmetric nonnegative matrix factorization. *Adv. Neural Inf. Process. Syst.* **31**, 5154–5164 (2018)
19. Nene, S.A., Nayar, S.K., Murase, H.: “Columbia Object Image Library” Tech. Rep. CUCS-005–96, Columbia University (1996)
20. Samaria, F.S., Harter, A.C.: Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994, Proceedings of the Second IEEE Workshop on*, pp. 138–142. IEEE (1994)
21. Graham, D., Allinson, N.: “Characterising virtual eigen signatures for general purpose face recognition,” Springer (1998)
22. Pan, J., Tong, Z., Li, H., Salzmann, M., Reid, I.: Deep subspace clustering network (2017)
23. Li, F., Qiao, H., Zhang, B.: Discriminatively boosted image clustering with fully convolutional auto-encoders. *Patt. Recogn.* **83**, 161–173 (2017)
24. Park, J., Lee, M., Chang, H.J., Lee, K., Jin, Y.C.: Symmetric graph convolutional autoencoder for unsupervised graph representation learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2020)

# Influence of Initialization and Modularization on the Performance of Network Morphism-Based Neural Architecture Search



Xuehui Chen , Xin Niu , Jingfei Jiang , Hengyue Pan , Peijie Dong , and Zimian Wei

**Abstract** Neural Architecture Search (NAS), the process of automatic network architecture design, has enabled remarkable progress over the last years on Computer Vision tasks. In this paper, we propose a novel and efficient NAS framework based on network morphism to further improve the performance of NAS algorithms. Firstly, we design four modular structures termed RBNC block, CBNR block, BNRC block and RCBN block which correspond to four initial neural network architectures and four modular network morphism methods. Each block is composed of a ReLU layer, a Batch-Norm layer and a convolutional layer. Then we introduce network morphism to correlate different modular structures for constructing network architectures. Moreover, we study the influence of different initial neural network architectures and modular network morphism methods on the performance of network morphism-based NAS algorithms through comparative experiments and ablation experiments. Finally, we find that the network morphism-based NAS algorithm that uses CBNR block for initialization and modularization is the best method to improve performance. Our proposed method achieves a test accuracy of 95.84% on CIFAR-10 with least parameters (only 2.72 M) and fewer search costs (2 GPU-days) for network architecture search.

**Keywords** Block-wise network morphism · Neural architecture search · Initialization · Modularization

## 1 Introduction

Neural Architecture Search, aiming at automatically designing network architectures by machines, has recently achieved a great success for many tasks, such as image classification [1, 2], object detection [3, 4] and semantic segmentation [5]. Given the search space, the NAS algorithm will search candidate networks with different search strategies. Then candidate networks are evaluated and by which a

---

X. Chen · X. Niu (✉) · J. Jiang · H. Pan · P. Dong · Z. Wei  
National University of Defense Technology, Changsha, China  
e-mail: [niuxin@nudt.edu.cn](mailto:niuxin@nudt.edu.cn)

best-performing neural architecture is selected. Recently, many methods have been proposed to perform NAS algorithms, mainly including gradient-based methods [6–8], reinforcement learning (RL) [2, 9, 10], and evolutionary methods [1, 11]. These NAS algorithms can achieve high performance, but they suffer from a large computational resource requirement. Moreover, each candidate network is trained from scratch, which is highly time-consuming. For instance, obtaining a SOTA architecture for CIFAR-10 require thousands of GPU-days by evolutionary algorithm [1] or reinforcement learning [10]. Searching for a smaller neural network with high performance in a short time is in urgent need.

Network morphism could be helpful for NAS algorithms by enabling a more efficient training [12]. Therefore, the NAS algorithm can use the weight inheritance method to accelerate the search process. Although the traditional network morphism method [13] is efficient, there is still some room for improvement. For example, the traditional layer-wise network morphism-based NAS algorithm in Auto-Keras [12] searches the best model on CIFAR-10, test accuracy of the best model is only 88.56%. In our proposed method, we modularize the large macro search space of NAS to improve the test accuracy. Specially, we regard the minimum block composed of convolutional, ReLU, and Batch-Norm layers as the basic unit, and study the influence of initialization and modularization on the performance of network morphism-based NAS algorithms. Experiments show that different initial neural network architectures will significantly affect the test accuracy of the best model obtained by NAS algorithms. Moreover, compared with the layer-wise network morphism method, the block-wise network morphism method can not only improve test accuracy of the obtained best model, but also search more neural network architectures and improve the search efficiency of NAS algorithms.

The main contributions of the paper are as follows:

- (1) We propose four simple modular structures, including ReLU-BatchNorm-Conv Block, Conv-BatchNorm-ReLU Block, BatchNorm-ReLU-Conv Block, and ReLU-Conv-BatchNorm Block, which correspond to four initial neural network architectures and four modular network morphism methods.
- (2) We design comparative experiments and ablation experiments for studying the influence of different initial neural network architectures on the performance of net-work morphism-based NAS algorithms.
- (3) We take the simple modular structure as the basic unit of network morphism operations for studying the influence of modularization on the performance of network morphism-based NAS algorithms.
- (4) We propose a novel and efficient NAS framework based on network morphism to further improve the performance of NAS algorithms. Our proposed method achieves a test accuracy of 95.84% on CIFAR-10 with least parameters (only 2.72 M) and fewer search costs (2 GPU-days) for network architecture search.

## 2 Related Work

### 2.1 Neural Architecture Search

In recent years, neural architecture search has aroused extensive attention. The principle of NAS is to automatically search the best neural network architecture for specific tasks through limited computing resources with human intervention as little as possible. The mathematical model of NAS is described as Eq. (1):

$$\left\{ \begin{array}{l} \arg \min_f = L(f, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{valid}}) \\ \text{s.t. } f \in \mathcal{F} \end{array} \right. \quad (1)$$

where  $\mathcal{F}$  denotes the search space of neural network architectures,  $L(\cdot)$  measures the loss of the architecture  $f$  on the validation dataset  $\mathcal{D}_{\text{valid}}$  after being trained on the training dataset  $\mathcal{D}_{\text{train}}$  [14].

Traditional NAS algorithms are highly time-consuming and require vast computing resources, which limit their application in real world. Therefore, network morphism [15, 16] and super-net are adopted by researchers to improve the search efficiency.

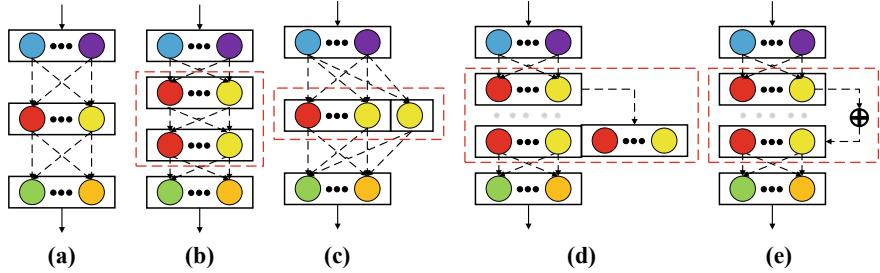
### 2.2 Traditional Layer-Wise Network Morphism Method

Network morphism methods can be categorized into layer-wise and block-wise. The layer-wise network morphism method takes the layer as the basic unit to morph the father network, while the block-wise network morphism method takes the block as the basic unit. The layer-wise network morphism operations mainly include “deepen network”, “widen network” and “add a skip connection” (see Fig. 1).

**Deepen network.** Randomly choose some position in the father network and insert a new layer which may be a ReLU layer, a Batch-Norm layer, or a convolutional layer. Specifically, when inserting the convolutional layer, the new convolutional layer initializes the weight matrix  $W^{i+1}$  as an identity matrix according to Net2DeeperNet [17] operation. when inserting the Batch-Norm layer, the offset and scale of the new Batch-Norm layer are initialized to the batch mean and batch variance to maintain the identity mapping.  $ReLU(x) = \max\{x, 0\}$  satisfies the constraint for activation function, so inserting ReLU layer is possible.

$$\forall x : \sigma(x) = \sigma(I\sigma(x)) \quad (2)$$

**Widen network.** Randomly choose some convolutional/dense layer in the father network. Increase the number of filters in the convolutional layer or the number of hidden cells in the dense layer to widen the neural network.



**Fig. 1** Visualization of layer-wise network morphism-based NAS algorithm. Given the father network **a**, **b** denotes the sub-network after depth morphing, **c** represents the sub-network after width morphing, **d** and **e** denote the sub-network after adding a skip connection by concatenation and addition. Different color circles represent different channels of a layer

**Add a skip connection.** Randomly choose two layers and add a new skip connection between the two layers either by concatenation or by addition. The parameters of  $W^{i+1}$  are initialized to 0 to maintain the identity mapping. The concatenation and addition operations are reformulated as:

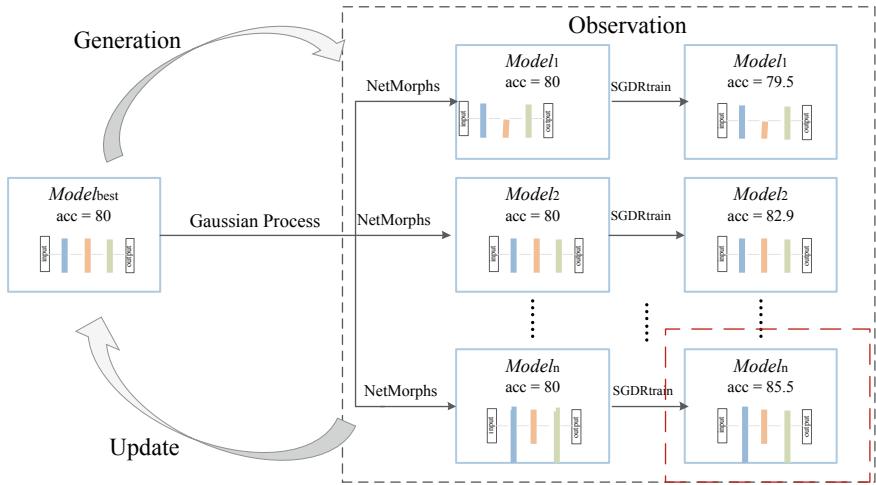
$$\begin{aligned} & \text{Concat}(\varphi(X^{i+1} \cdot W^{i+1}), \varphi(X^j)) \\ & \text{Add}(\varphi(X^j), \varphi(X^{i+1} \cdot W^{i+1})) \end{aligned} \quad (3)$$

### 2.3 Efficient Network Morphism-Based NAS Algorithm

Many works [15, 16] have used network morphism to accelerate the training and search process of NAS. For example, Auto-Keras [12] regards the NAS problem as a black-box optimization task and uses Bayesian optimization (BO) algorithm to guide the network morphism for the optimal solution of the black-box target (see Fig. 2). The BO algorithm iteratively conducts: (1) Update: train the Gaussian Process model with the existing architectures and their accuracy; (2) Generation: generate the next sub-network architecture to observe by the acquisition function; (3) Observation: obtain the ground-truth accuracy of the generated neural architecture by training. [12]

## 3 Proposed Methods

The critical idea of network morphism-based NAS algorithms is to make the initial neural network architecture under the guidance of BO algorithm to select proper network morphism operations. After multiple network morphism operations, the



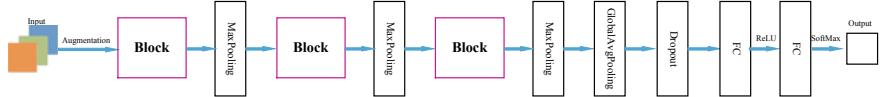
**Fig. 2** Visualization of the efficient network morphism-based NAS algorithm. Different color rectangles represent different layers and the length of rectangles represents the width of layers

network with the highest accuracy will be taken as the best model. The selection of the initial model and network morphism method influence the performance of NAS algorithms. Therefore, we study the influence of initialization and modularization on the performance of network morphism-based NAS algorithms from initial neural network architectures and modular network morphism methods in the section.

### 3.1 Design of Initial Neural Network Architecture

In recent years, typical manually designed architectures, such as ResNet [18], DenseNet [19], MobileNet [20] and so on, have greatly improved the accuracy of image classification. These architectures with high performance are often stacked by blocks. For example, ResNet is stacked by Conv-BatchNorm-ReLU block as the basic unit, and DenseNet is stacked by BatchNorm-ReLU-Conv block. The initial model in Auto-Keras [12], is initialized with the three-layer CNN. Each convolutional layer is a convolutional block composed of a ReLU layer, a Batch-Norm layer, a convolutional layer, and a pooling layer. Following the design principle of the initial model in Auto-Keras, we define the initial model architecture stacked by blocks (see Fig. 3). It starts with three convolutional blocks and MaxPooling layers, which are connected alternately. Then a GlobalAveragePooling layer and a dropout layer are added, followed by two dense layers and a Softmax layer.

In addition, we follow the modular characteristics in typical manually designed neural network architectures, design four modular structures as a convolutional



**Fig. 3** The initial model

block in the initial model, including **ReLU-BatchNorm-Conv block** (RBNC), **Conv-BatchNorm-ReLU block** (CBNR), **BatchNorm-ReLU-Conv block** (BNRC), and **ReLU-Conv-BatchNorm block** (RCBN).

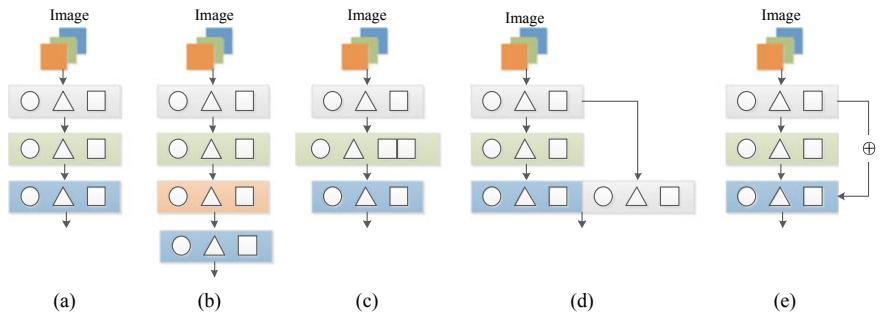
### 3.2 Design of Block-Wise Network Morphism Methods

Block-wise network morphism methods morph the father network by convolutional block (Sect. 3.1) as the basic unit, which is composed of a ReLU layer, a Batch-Norm layer, and a convolutional layer. Block-wise Network morphism operations include three operations (see Fig. 4).

**Insert a block.** Randomly select a location in the network and insert a new block. We initialize the convolution layer and Batch-Norm layer same as layer-wise network morphism methods to maintain identity mapping. Convolution kernel size is randomly selected, and the number of channels is consistent with neighbor convolution layers.

**Increase the number of filters of a convolutional layer.** Randomly select a block in the network and proportionally widen channel numbers of the convolutional block.

**Add a skip connection.** Randomly select two blocks in the network and add a new skip connection between them either by concatenation or by addition.



**Fig. 4** Visualization of block-wise network morphism-based NAS algorithm. Given the father network **a**, **b** denotes the sub-network after inserting a new block, **c** represents the sub-network after increasing the number of filters of a convolution layer, **d** and **e** denote the sub-network after adding a skip connection by concatenation and addition. Different color rectangles represent different layers. A circle/triangle/square represents a ReLU/Batch-Norm/ convolutional layer

## 4 Experiments and Results

By comparative experiments, we study the influence of initialization and modularization on the performance of network morphism-based NAS algorithms. Considering that all block-wise network morphism operations can solely increase a network’s size, the neural network architecture will become more complex. Therefore, we set all experiments to run on 4 NVIDIA 2080Ti GPUs for only 12 h.

### 4.1 Baseline Experiment

**Dataset.** The CIFAR-10 dataset contains ten kinds of object color pictures. Each category includes 6000  $32 \times 32$  pixels images, 60,000 images in total, 50,000 of which are training datasets, and 10,000 are test datasets.

**Initial Model.** The initial neural network architecture uses RBNC block for initialization. All convolutional layers contain 64 filters,  $kernel\_size = 3 \times 3$ . Each MaxPooling layer has a stride of two and  $keep\_prob$  of the dropout layer is equal to 0.75.

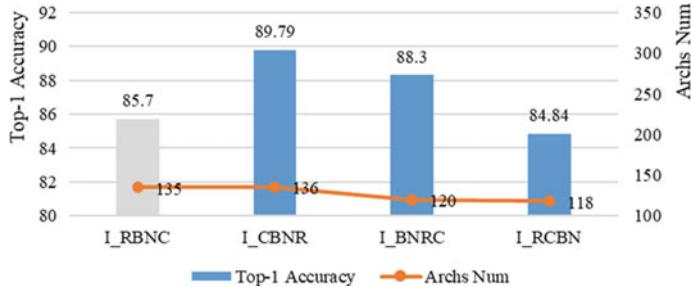
**Search on CIFAR-10.** We use the layer-wise network morphism method (Sect. 2.3) to search neural network architectures. The  $kernel\_size$  of new convolutional layers is equal to  $1 \times 1$ ,  $3 \times 3$ , or  $5 \times 5$ ,  $stride = 1$ . All the generated architectures are trained 200 epochs by using SGD optimizer with learning rate = 0.001, momentum = 0.9, weight decay = 0.00001, batch size = 128. We also use the early stop method to optimize neural networks.

**Post-Training of the Best Neural Architecture Obtained.** We use the AutoAugment [21] method for preprocessing. The model is trained on the training dataset until convergence using Cutout [22] and Mixup [23] (Cutout size of  $16 \times 16$  and  $\alpha = 1$  for Mixup).

The obtained best model is trained by using SGD with batch size = 128 and learning rate  $l = 0.1$  for 50 epochs to accelerate the convergence process. Then we used SGDR with initial learning rate  $l_{max} = 0.1$ ,  $T_0 = 1$ , and  $T_{multi} = 2$  for 600 epochs. Finally, the accuracy on the test dataset is 85.7%.

### 4.2 Influence of Different Initial Neural Network Architectures on the Performance of Network Morphism-Based NAS Algorithms

To compare the influence of different initial neural network architectures on the performance of network morphism-based NAS algorithms, we design a group of comparative experiments, statistically analyze test accuracy of the best model and number of neural networks searched by NAS algorithm within 12 h. Then we compare



**Fig. 5** Performances of layer-wise network morphism-based NAS algorithms with different initial neural network architectures. Gray rectangle denotes the result of the baseline experiment, and  $I\_*$  denotes the initial neural network architecture with \* block. Tag \* represents RBNC/CBNR/BNRC/RCBN

the performance of NAS algorithm from two aspects of test accuracy and search efficiency. Dataset and Post-training method are the same as those in the baseline experiment. All experiments use the layer-wise network morphism-based NAS algorithm. The blocks used in the initial model (Fig. 1) of comparative experiments are CBNR block, BNRC block, and RCBN block, respectively. Results of comparative experiments are as shown in Fig. 5.

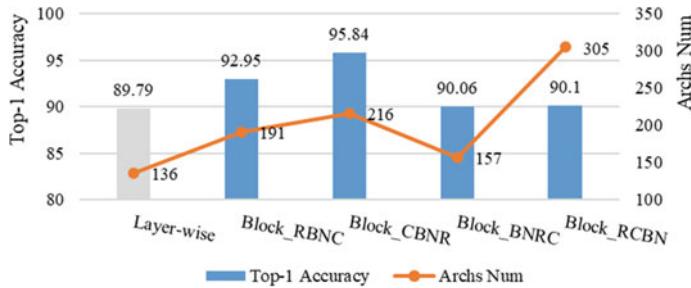
Compared with the baseline experiment, the layer-wise network morphism-based NAS algorithm that adopts the initial model with CBNR block or BNRC block, can achieve better performance. Test accuracy is respectively improved by 4.09% and 2.6%. The performance of the layer-wise network morphism-based NAS algorithm that adopts the initial model with RCBN block is reduced, and test accuracy is reduced by 0.86%. The number of architectures searched by NAS algorithm within 12 h is in the range of 100–150. Experimental results show that different initial neural network architectures have significant impact on the performance of network morphism-based NAS algorithms. Particularly, choosing the right initial model can improve the accuracy of obtaining the optimal model.

#### 4.3 Influence of Modular Network Morphism Methods on the Performance of Network Morphism-Based NAS Algorithms

As shown in Fig. 5, the layer-wise network morphism-based NAS algorithm that adopts the initial model with CBNR block has the highest performance. Therefore, we use it (accuracy of the best model is 89.79%, the number of architectures searched by NAS algorithm within 12 h is 136) as baseline. Moreover, all comparative experiments adopt the same initial model. For comparing the influence of different modular network morphism methods on the performance of network morphism-based NAS algorithms, RBNC block, CBNR block, BNRC block, and RCBN block

are used as the basic unit of block-wise network morphism operation in comparative experiments. Results are as shown in Fig. 6.

In terms of test accuracy, there is no apparent difference between baseline and block-wise network morphism-based NAS algorithm with BNRC block, and RCBN block. Moreover, compared with baseline, the block-wise network morphism-based NAS algorithm with RBNC block or CBNR block can achieve better performance. Test accuracy is respectively improved by 3.16% and 6.05%. Remarkably, the NAS algorithm that uses CBNR block for initialization and modularization can achieve the highest performance (test accuracy achieves 95.84%). The comparison against state-of-the-art recognition results on CIFAR-10 is presented in Table 1. Our method can design an effective neural network architecture with the least parameters and



**Fig. 6** Performances of block-wise network morphism-based NAS algorithms with different modular network morphism methods. Gray rectangle denotes baseline. Block\_ \* denotes the block-wise network morphism-based NAS algorithm with \* block. Tag \* represents RBNC/CBNR/BNRC/RCBN

**Table 1** Comparison with results of automatically designed architectures by SOTA NAS methods on CIFAR-10

Method	Params (Mil.)	Search time (GPU-days)	Test accuracy (%)
AmoebaNet-A [1]	3.2	3150	96.66
Large-scale Evolution [24]†	5.4	2600	94.6
NAS-v3 [10]	37.4	1800	96.35
NASNet-A [2]	3.3	1800	97.35
Hierarchical Evolution [25]†	15.7	300	96.25
PNAS [26]†	3.2	225	96.59
NAONet [27]	128	200	97.89
EAS [28]†	23.4	10	95.77
DARTS [7]	3.4	4	97.17
ENAS [11]	4.6	0.45	97.11
Auto-Keras[12] †	–	0.5	88.56
Ours	<b>2.72</b>	2	95.84

Results marked with † are not trained with Cutout [22]

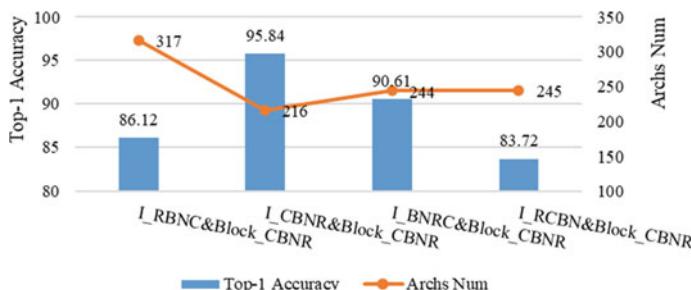
FLOPs. Experimental results show that modular network morphism methods have significant impact on the performance of network morphism-based NAS algorithms.

In terms of search efficiency, compared with the layer-wise network morphism-based NAS algorithm, the number of architectures searched by the block-wise network morphism-based NAS algorithm with CBNR block or RCBN block within 12 h is in the range of 200 to 350. The search efficiency is doubled. For the block-wise network morphism-based NAS algorithms with RBNC block and BNRC block, the difference in search efficiency is not apparent. Experiments show that modular network morphism methods have significant impact on the performance of network morphism-based NAS algorithms. Particularly, choosing the right mode network morphism method can not only improve the accuracy of obtaining the optimal model, but also speed up the search efficiency.

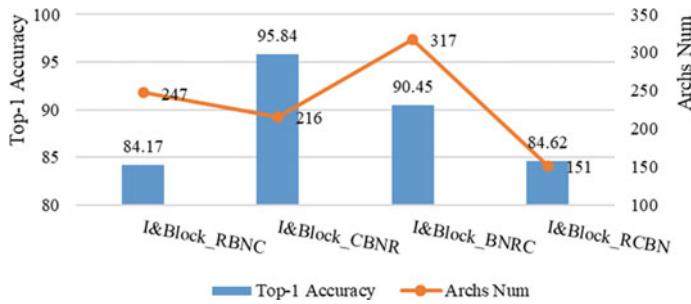
#### 4.4 Ablation Experiment

To ensure the completeness of experiments, we study the influence of different initial neural network architectures on the performance of block-wise network morphism-based NAS algorithm with CBNR block. It can be seen from Figs. 5 and 7 that, the influence of different initial neural network architectures on the performance of both layer-wise and block-wise network morphism-based NAS algorithms is consistent. The performance of the network morphism-based NAS algorithm that adopts the initial model with RBNC block or RCBN block is relatively low. The network morphism-based NAS algorithm that adopts the initial model with CBNR block or BNRC block can achieve better performance. Remarkably, the network morphism-based NAS algorithm that adopts the initial model with CBNR block has the highest performance.

Considering the block consistency, we design experiments that adopt the same block in both the initial model and the modular network morphism method.



**Fig. 7** Performances of block-wise network morphism-based NAS algorithms with different initial neural network architectures.  $L_{*}\&Block\_CBNR$  denotes the network morphism-based NAS algorithm that adopts the initial model with \* block and block-wise network morphism method with CBNR block. Tag \* represents RBNC/CBNR/BNRC/RCBN



**Fig. 8** Performances of network morphism-based NAS algorithms with different blocks for initialization and modularization. I&Block\_\* denotes the network morphism-based NAS algorithm with \* block for initialization and modularization. Tag \* represents RBNC/CBNR/ BNRC/RCBN

Then we analyze the influence of initialization and modularization on the performance of network morphism-based NAS algorithms. As shown in Fig. 8, the network morphism-based NAS algorithm that uses CBNR block for initialization and modularization still has the highest performance.

## 5 Conclusion

In this paper, we study the influence of initialization and modularization on the performance of network morphism-based NAS algorithms. We finally propose a novel and efficient NAS framework that uses CBNR block for initialization and modularization. This method searched the best-performing network architecture with 2.72 M parameters and the architecture achieves a test accuracy of 95.84% on Cifar-10. The search cost of our proposed method is less than 12 h on 4 NVIDIA 2080Ti GPUs. Experimental results reveal that modular design is a critical component in network morphism-based NAS algorithms.

**Acknowledgements** We acknowledge support from the Parallel and Distributed Processing Laboratory of National University of Defense Technology. This work was supported by National Key Research & Development Program of P.R. of China (2018YFB1003400).

## References

1. Real, E., Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: AAAI. pp. 4780–4789. AAAI Press (2019)
2. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: CVPR. pp. 8697–8710. IEEE Computer Society (2018)
3. Chen, Y., Yang, T., Zhang, X., Meng, G., Pan, C., Sun, J.: Detnas: Neural architecture search on object detection. CoRR, abs/1903.10979 (2019)

4. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: CVPR. pp. 7036–7045. Computer Vision Foundation / IEEE (2019)
5. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)
6. Cai, H., Zhu, L., Han, S.: Proxylessnas: Direct neural architecture search on target task and hardware. In: ICLR (Poster). OpenReview.net (2019)
7. Liu, H., Simonyan, K., Yang, Y.: DARTS: differentiable architecture search. CoRR, abs/1806.09055 (2018)
8. Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., Keutzer, K.: Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In: CVPR. pp. 10734–10742. Computer Vision Foundation / IEEE (2019)
9. Pham, H., Guan, M.Y., Zoph, B., Le, Q.V., Dean, J.: Efficient neural architecture search via parameter sharing. In: ICML. Proceedings of Machine Learning Research, vol. 80, pp. 4092–4101. PMLR (2018)
10. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. In: ICLR. OpenReview.net (2017)
11. Desell, T.: Large scale evolution of convolutional neural networks using volunteer computing. In: Bosman, P.A.N. (ed.) GECCO (Companion). pp. 127–128. ACM (2017)
12. Jin, H., Song, Q., Hu, X.: Auto-keras: An efficient neural architecture search system. In: Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., Karypis, G. (eds.) KDD. pp. 1946–1956. ACM (2019)
13. Wei, T., Wang, C., Rui, Y., Chen, C.W.: Network morphism. In: Balcan, M.F., Weinberger, K.Q. (eds.) ICML. JMLR Workshop and Conference Proceedings, vol. 48, pp. 564–572. JMLR.org (2016)
14. Liu, Y., Sun, Y., Xue, B., Zhang, M., Yen, G.G.: A survey on evolutionary neural architecture search. CoRR, abs/2008.10937 (2020)
15. Elsken, T., Metzen, J.H., Hutter, F.: Simple and efficient architecture search for convolutional neural networks. In: ICLR (Workshop). OpenReview.net (2018)
16. Wei, T., Wang, C., Chen, C.W.: Stable network morphism. In: IJCNN. pp. 1–8. IEEE (2019)
17. Chen, T., Goodfellow, I.J., Shlens, J.: Net2net: Accelerating learning via knowledge transfer. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2016)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
19. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017)
20. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017)
21. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. CoRR, abs/1805.09501, (2018).
22. Devries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. CoRR, abs/1708.04552 (2017)
23. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. CoRR, abs/1710.09412 (2017).
24. Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y.L., Tan, J., Le, Q., Kurakin, A.: Large-scale evolution of image classifiers (2017)
25. Liu, H., Simonyan, K., Vinyals, O., Fernando, C., Kavukcuoglu, K.: Hierarchical representations for efficient architecture search. In: ICLR (Poster). OpenReview.net (2018)
26. Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.J., Fei-Fei, L., Yuille, A.L., Huang, J., Murphy, K.: Progressive neural architecture search. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV. Lecture Notes in Computer Science, vol. 11205, pp. 19–35. Springer (2018)

27. Luo, R., Tian, F., Qin, T., Chen, E., Liu, T.Y.: Neural architecture optimization. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) NeurIPS. pp. 7827–7838 (2018)
28. Cai, H., Chen, T., Zhang, W., Yu, Y., Wang, J.: Efficient architecture search by network transformation. In: McIlraith, S.A., Weinberger, K.Q. (eds.) AAAI. pp. 2787–2794. AAAI Press (2018)

# A Document Image Quality Assessment Method Based on Feature Fusion



Weisheng Wang , Zhiyang Yan , and Hongli Lin

**Abstract** Document image quality assessment (DIQA) is an essential step in the development of optical character recognition (OCR) products. Due to the complex and diverse distortion types in the real captured document images, DIQA is still a challenging problem. In this paper, we propose a new DIQA model, which is based on the feature fusion in convolutional neural network (CNN). In our network, shallow network part is used to extract low-level local features of document images to represent local non-uniform distortions. And deep network part is used to learn global features to represent global uniform distortions in document images. In addition, a quality regression network is used to predict the document image quality score by using the fusion of the low-level and deep-level features. Experimental results demonstrate that our model outperforms the state-of-the-art methods on complex distortion datasets.

**Keywords** Document image quality assessment · Document image · DIQA · Feature fusion

## 1 Introduction

As the popularity of smart devices grows, document image recognition is not just for traditional scanned text, but more for real document images captured by smart device cameras. In recent years, many Internet companies have developed document image recognition services, of which OCR services occupy the mainstream position. The performance of the OCR engine is closely related to the quality of the document image, however, due to the defects of the shooting equipment or photography skills, the document image will be distorted during the capture process, resulting in different degrees of image quality problems [18] and lower OCR accuracy. In this case, the important information in the document image is recognized incorrectly or lost, causing immeasurable costs. Therefore, it makes sense to apply DIQA before

---

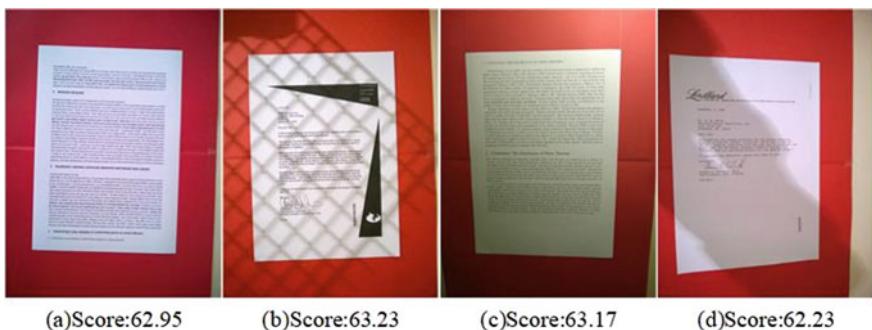
W. Wang · Z. Yan · H. Lin ()

College of Computer Science and Electronic Engineering, Hunan University, Hunan, China  
e-mail: [hllin@hnu.edu.cn](mailto:hllin@hnu.edu.cn)

recognizing the document image, so as to remove the low-quality image, or to further restore [15] or enhance [16] the document image.

Generally, the field of image quality assessment is divided into the quality assessment of natural scene images and document images. In recent years, the field of natural scene image quality assessment has developed rapidly [17]. However, there are not only significant difference between document images and natural scene images in terms of structure and measurement formulas [6], but the goals of the two are also very different. Unlike natural images quality assessment, which can be evaluated based on human perception, DIQA can be evaluated based on OCR accuracy. Consequently, the natural scene image quality evaluation model may not be directly applied to document images [18].

In the past few years, many people have made great efforts to assess the quality of document images through different methods. Although some progresses have been achieved, there are still huge challenges to the evaluation of document image quality with complex multiple distortions. There are various types of document image distortions, and multiple distortions may be concentrated on one image in unexpected ways [13]. Because of the diversity of distortion types, different document images have different types of distortion, but they may have similar OCR accuracy, as shown in Fig. 1. Nevertheless, In the existing document image quality assessment methods, whether based on traditional manual features [5, 7, 14], or learning based document image quality assessment methods [4, 8, 9, 11], they either only tend to extract low-level or global features of the image, ignoring deep-seated features and local features, yet, in the document image captured by smart device camera, there may be global distortion caused by defocus or illumination, or local distortion caused by lens jitter or shadow. As a result, the algorithms which only tend to extract low level features and the algorithms which only tend to learn global features with deep models still have not worked well. Therefore, aggregating both local distortion features and global distortion features, and then predicting document image quality upon this multi-scale representation is an efficient approach.



**Fig. 1** **a, b, c** and **d** are four document images with different degrees of distortion on the SmartDoc-QA [13] data set. Different distortion features are mapped to similar OCR accuracy in these four images

In this paper, we develop a DIQA model for the complex distortion of real document images. We extract low-level local features and deep-level global features from multi-scale feature maps, which are fused to deal with the distortion of diversity.

Our model uses two network structures to extract low-level and deep-level features, and then merges local distortions features which are captured by a local feature extractor with global quality features. A final quality score is predicted through a quality regression network, which is trained by fusing low-level and deep-level features. Conducted a series of experiments demonstrate that our model achieves significant effects on complex datasets in terms of DIQA and precedes the latest DIQA methods [4, 8, 9, 11] reported in the literature.

The following chapters of this paper will describe the related work, our specific methods, the analysis of experimental results and the summary of this paper.

## 2 Related Work

In the process of OCR, the distorted document image may lose key information, resulting in incorrect recognition results. As a consequence, it is very significant to add DIQA in OCR process. Given that DIQA is bound up with OCR accuracy, OCR accuracy is adopted as the quality descriptor in most DIQA methods. The current latest DIQA methods are usually divided into two categories: metric-based methods and learning-based methods.

### 2.1 *The Metric-Based DIQA Methods*

The metric-based DIQA method generally extracts different manual features to generate a quality map to the quality score of the document image. Kumar et al. [5] used the grayscale change of image after median filtering to calculate the sharpness information to assess the image quality. In [14], Nayef et al. developed an DIQA method based on OCR accuracy. This method calculates the quality score through proportional weighted summation based on the dependence between different distortions of the document image and combined with a specific distortion measure. In Kumar et al. [7], the quality score is calculated by character gradient. However, these techniques focus only on the specific characteristics of the image, and the effect is not obvious for document images with complex and diverse distortion types.

### 2.2 *The Learning-Based DIQA Methods*

The learning-based document image quality assessment model generally includes two steps: feature extraction and quality score regression. In recent years, Kang et al.

[4] proposed a CNN model to assess document image quality. Li et al. [9] implemented an attention-based recurrent neural network (RNN) for DIQA. Different from the traditional DIQA method based on OCR accuracy, this framework integrates CNN and RNN to form a glimpse-RNN-Action combined network. Lu et al. [11] applied the deep transfer learning method to DIQA and put forward a deep CNN model. In [8], Li et al. proposed a DIQA framework where the overall quality score was weighted by the quality score of each text block. Nevertheless, when the document image quality is low, these methods will lose some key text information or text lines that affect the accuracy of OCR, resulting in inaccurate quality prediction.

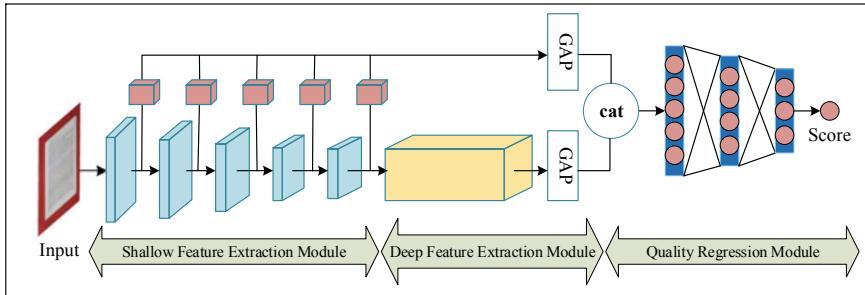
Although these methods have made great progress, these methods are more inclined to extract the low-level features of the text area, ignoring the deeper and complex semantic features contained in the uniform distorted document image. In practical applications, a truly distorted document image may be merged by multiple distortions in a complicated manner, and low-level features cannot fully represent the document image with diverse distortions. In addition, diverse distortions may exist locally or globally, and the sensitivity of the OCR engine is determined by these two conditions. In this paper, inspired by [17], we built a new DIQA framework. Our framework combines low-level features and deep features while fusing local non-uniform distortions and global uniform distortions, and the results are obtained through a quality regression module. The experimental results prove that our quality evaluation model advantage over ones reported in the literature.

### 3 Method

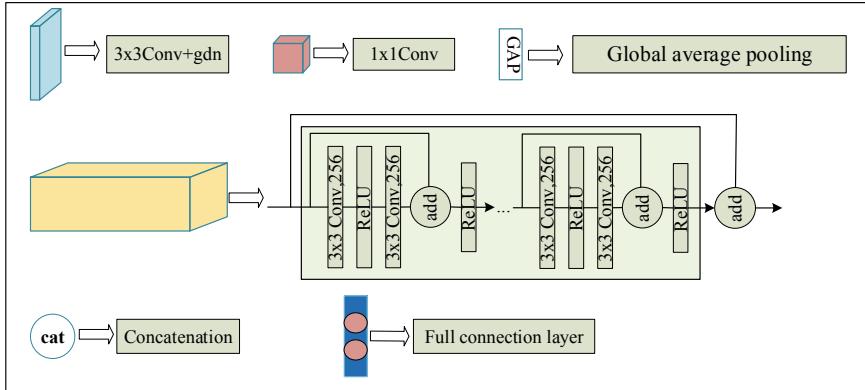
In this section, we develop a DIQA network, and the network architecture is shown in Fig. 2, including shallow feature extraction module, deep feature extraction module and quality regression module. Each module is described in detail below.

#### 3.1 Shallow Feature Extraction Module

The main distortion problems of document images include: illumination, blur, scene background, stains, and color degradation, resolution, etc., resulting in image quality problems [13]. Therefore, we try to keep the quality information of the original image in the low-level feature extraction stage, and perform preliminary extraction of the quality information. In this part, we are inspired by [12] and combine  $3 \times 3$  convolution and generalized divisive normalization (GDN) [2] as the backbone, where GDN is highly non-linear [1] and has spatial adaptability. In order to better capture the local distortion information, we use  $1 \times 1$  convolution and global average pooling (GAP) to convert multi-scale features into local feature vectors. It was proved in [17] that this structure can be considered as an attention-based local feature extractor,



Each component is described below:



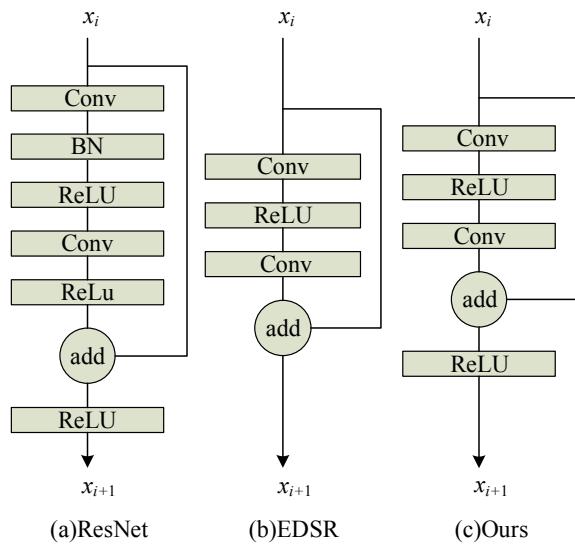
**Fig. 2** The DIQA model framework proposed in this paper contains three modules: shallow feature extraction module, deep feature extraction module and quality regression module

which can perceive the regional features corresponding to the local distortion, so as to better capture its quality.

### 3.2 Deep Feature Extraction Module

Because of the diversity and complexity of document image distortion types, it is necessary to extract deeper quality features from document images. For fear of preventing the degradation caused by the increase of the network depth, we use the structure shown in Fig. 3 to increase the network depth. In Fig. 3c, we removed the batch normalization (BN) layer in ResNet [3], which not only reduces network consumption, but also increases the flexibility of the network. Among them, we set the size of the convolution kernel to 3 and the number of channels to 256. After removal, we can stack more network layers, and each layer can extract more features [10]. We extract deep semantic features by stacking 18 structures in Fig. 3c, and merge the shallow semantic features extracted in Sect. 3.1 with the output of the

**Fig. 3** Comparison of the residual structure of ResNet [3], EDSR [10] and this paper



deep semantic feature module. Finally, GAP is also fused with local features to feed the quality regression network.

### 3.3 Quality Regression Module

In this part, our goal is to map the previously extracted image features to the quality score, so we build a quality regression network with three fully connected layers. As shown in Fig. 2, the multi-scale feature fusion vector is used as input and propagated through three fully connected layers where the ReLU function serves as the activation function, and finally the document image quality score is obtained. Our model can be described as

$$s = \varphi(L(x), D(x), \gamma) \quad (1)$$

where  $\varphi$  represents the network model,  $x$  and  $L(x)$  are the input image and the output of the shallow feature extraction network respectively,  $D(x)$  is recorded as the result of the deep feature extraction network, and  $\gamma$  is the model parameter.

### 3.4 Implementation Details

In our experiment, the dataset is spilt into a training set and a test set in a ratio of 8–2, and set the batch size to 16. Adam optimizer is selected to optimize the prediction network, where the learning rate and the betas are set to 1e-4 and (0.9, 0.999) respectively, meanwhile, eps is adjusted to 1e-8, and weight decay is 0. In the training phase, we randomly cut the input image into  $224 \times 224 \times 3$  to form 16 patches, and the ground truth of each patch is the same as the input image. For the entire training process, the loss function is  $l_1$ -norm.

$$\ell = \|s - \hat{s}\|_1 = \sum_{i=1}^n |s_i - \hat{s}_i| \quad (2)$$

where  $s_i$  and  $\hat{s}_i$  represent the ground truth and predicted quality score of the  $i$ -th patch,  $n$  is the total number of patches. In the testing phase, the sample is also randomly divided into 16 patches, and the quality scores of the 16 patches are averaged to obtain the final quality score.

## 4 Experiment

Our model is evaluated on two public datasets Sharpness-OCR-Correlation (SOC) [6] and SmartDoc-QA [13] and compared them with the state-of-the-art approaches.

### 4.1 Datasets and Evaluation Metrics

The SOC dataset is made up of 175 document images with a resolution of  $1840 \times 3264$  and is composed of 25 English documents taken with a smartphone, and each takes 6–8 images with different focal lengths to produce varying degrees of distortion. The SOC dataset uses three OCR engines (ABBY FineReader, Tesseract, and Omnipage) to evaluate the OCR accuracy of each image. In our experiments, we use the mean results of the three OCR engines as the ground truth. The SmartDoc-QA dataset is a more complex data set with more distortion types. The dataset contains 4260 document images, which were taken from 30 documents by two different mobile phones. The 30 document images are mainly composed of three types of official documents, old official documents and receipts. The OCR accuracy of this dataset is the recognition results of the FineReader and Tesseract OCR engines. Similarly, we calculate the mean of the two OCR recognition results as the ground truth.

We choose two conventional evaluation indicators: Spearman Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC)

to evaluate the performance of the model. SROCC indicates the monotony of the predicted results and is defined as

$$SROCC = 1 - \frac{6 \sum_i^n d_i^2}{n(n^2 - 1)} \quad (3)$$

where  $d_i$  is the rank difference between the prediction result of the  $i$ -th test image and the ground truth,  $n$  is the number of test set. PLCC is commonly used to describe the accuracy of prediction results and is defined as

$$PLCC = \frac{\sum_i^n (s_i - s_m)(\hat{s}_i - \hat{s}_m)}{\sqrt{\sum_i^n (s_i - s_m)^2 \sum_i^n (\hat{s}_i - \hat{s}_m)^2}} \quad (4)$$

where  $s_i$  and  $\hat{s}_i$  are the ground truth and prediction result of the  $i$ -th test image separately,  $s_m$  and  $\hat{s}_m$  are the mean values of all ground truth and predictions,  $n$  is the number of test images. The larger the value of these two indicators, the better the performance, and the range is between 0 and 1.

## 4.2 Comparison with the State-Of-The-Art Methods

We have compared seven latest DIQA methods, including three metric-based methods: Sharpness [5], MetricNR [14] and CG-DIQA [7], four learning-based methods: CNN [4], RNN [9], TL [11] and DTL [8]. As shown in Table 1, on the SOC dataset, our method is significantly better than other methods. Our PLCC results are

**Table 1** Comparison of PLCC and SROCC results on SOC and SmartDoc-QA datasets with the latest methods

Methods	SOC		SmartDoc-QA	
	PLCC	SROCC	PLCC	SROCC
Sharpness [5]	N/A	N/A	0.624	0.596
MetricNR [14]	0.887	0.820	N/A	N/A
CG-DIQA [7]	0.906	0.856	0.625	0.631
CNN [4]	0.950	0.898	N/A	N/A
RNN [9]	0.956	0.916	0.814	<b>0.865</b>
TL [11]	0.914	0.872	0.743	0.757
DTL [8]	0.965	0.931	N/A	N/A
Ours	<b>0.991</b>	<b>0.968</b>	<b>0.956</b>	0.854

The bold values represent the optimal results of all the DIQA methods that were compared

leading in the other four methods and the SROCC are only slightly lower than RNN [9] for SmartDoc-QA dataset. Among them, the result of DTL on the data set SOC is better than the other methods. Smartdoc-QA dataset is more complex and has more types of distortion. our method performs on this dataset is slightly lower than that on the SOC dataset. This is because 40% of the 2160 document images scored by the Tesseract OCR engine on the Smartdoc-QA dataset have a result of 0%, which means that the OCR accuracy distribution of this dataset is unbalanced. From the results of PLCC, our method is still superior to the four most advanced methods on Smartdoc-QA dataset. In the results of SROCC, the attention mechanism-based RNN model [9] is better than the other four approaches, and also slightly exceed our method, which shows the attention-based RNN model [9] has better results on the monotonicity of prediction. In addition, our method is greatly superior to the other three methods in SROCC results. From the discussion above, it is obvious that our approach has an excellent performance for DIQA.

### 4.3 Ablation Study

We performed ablation experiments on the SOC datasets and the SmartDoc-QA dataset to assess the effectiveness of each components in our DIQA framework. We first proved the effectiveness of low-level feature extraction network (LC) and deep- level feature extraction network (DC). The results are shown in Table 2. Both indicators are superior to all current technologies on SOC datasets, and PLCC results are significantly superior to other methods on SmartDoc-QA dataset. Then we verify the effectiveness of the local distortion feature extraction module (MS). When LC is added to the local distortion feature extraction module, LC improves on both datasets. It is significantly improved by 1.3% on the SmartDoc-QA dataset in SROCC. And when we Combining LC, MS and DC, our model has been further improved in SROCC and PLCC, which reached 96.8% and 99.1% on the SOC dataset, and 85.4% and 95.6% on the SmartDoc-QA dataset.

**Table 2** Results of ablation experiment on SOC and SmartDoc-QA datasets

Components	SOC		SmartDoc-QA	
	PLCC	SROCC	PLCC	SROCC
LC	0.985	0.964	0.944	0.835
LC + MS	0.986	0.967	0.946	0.848
DC	0.969	0.955	0.952	0.837
LC + MS + DC	<b>0.991</b>	<b>0.968</b>	<b>0.956</b>	<b>0.854</b>

## 5 Conclusion

This paper proposes a new CNN model based on feature fusion to evaluate document image quality. Our model takes account of the diverse, local and global distortions of real document images by feature fusion, rather than the distortions in single aspect. In order to better predict the quality of real distorted document images, our shallow feature extraction module extracts low-level quality information and local distortion features, and try to preserve the original quality of image. Then we use the deep feature extraction module to acquire the high-level information of the distortion features, and finally combine the two features while fusing the local distortion features with the global semantics, and feed them to the quality regression module to get the final quality score. The experimental results prove that our model shows strong robustness to both simple distortion and complex multiple distortion document images.

In addition, this method explores the DIQA method through feature fusion, and also provides a prospect for multiple distortion document image quality evaluation in the field of document image quality evaluation in the future.

## References

1. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-End Optimized Image Compression. ArXiv abs/1611.01704 (2017)
2. Ballé, J., Laparra, V., Simoncelli, E.P.: Density modeling of images using a generalized normalization transformation. Int. Conf. Learning Represent. (2015)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>, <https://doi.org/10.1109/CVPR.90>
4. Kang, L., Ye, P., Li, Y., Doermann, D.: A deep learning approach to document image quality assessment. In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 2570–2574 (2014)
5. Kumar, J., Chen, F., Doermann, D.: Sharpness estimation for document and scene images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 3292–3295 (2012)
6. Kumar, J., Ye, P., Doermann, D.S.: A dataset for quality assessment of cam-era captured document images. In: Iwamura, M., Shafait, F. (eds.) Camera-Based Document Analysis and Recognition - 5th International Workshop, CB-DAR 2013, Washington, DC, USA, August 23, 2013, Revised Selected Papers. Lecture Notes in Computer Science, vol. 8357, pp. 113–125. Springer (2013)
7. Li, H., Zhu, F., Qiu, J.: CG-DIQA: No-reference Document Image Quality Assessment Based on Character Gradient. In: 2018 24<sup>th</sup> International Conference on Pattern Recognition (ICPR). pp. 3622–3626 (2018)
8. Li, H., Zhu, F., Qiu, J.: Towards Document Image Quality Assessment: A Text Line Based Framework and a Synthetic Text Line Image Dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 551–558 (2019)
9. Li, P., Peng, L., Cai, J., Ding, X., Ge, S.: Attention based RNN Model for Document Image Quality Assessment. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 819–825 (2017)

10. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
11. Lu, T., Dooms, A.: A Deep Transfer Learning Approach to Document Image Quality Assessment. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1372–1377 (2019)
12. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-End blind image quality assessment using deep neural networks. *IEEE Trans. Image Process.* **27**(3), 1202–1213 (2018)
13. Nayef, N., Luqman, M.M., Prum, S., Eskenazi, S., Chazalon, J., Ogier, J.: SmartDoc-QA: A dataset for quality assessment of smartphone captured document images - single and multiple distortions. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1231–1235 (2015)
14. Nayef, N., Ogier, J.: Metric-based no-reference quality assessment of heterogeneous document images. In: Ringger, E.K., Lamirov, B. (eds.) Document Recognition and Retrieval XXII, San Francisco, California, USA, February 11–12, 2015. SPIE Proceedings, vol. 9402, pp. 94020L. SPIE (2015)
15. Ouafek, N., Kholladi, M.: A binarization method for degraded document image using artificial neural network and interpolation inpainting. In: 2018 4th International Conference on Optimization and Applications (ICOA). pp. 1–5 (2018)
16. Sharma, P., Sharma, S.: An analysis of vision based techniques for quality assessment and enhancement of camera captured document images. In: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). pp. 425–428 (2016)
17. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3664–3673 (2020)
18. Ye, P., Doermann, D.: Document Image Quality Assessment: A Brief Survey. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 723–727 (Aug 2013)

# Design of Simulation Device for Greenhouse Control



Yunsong Jia , Shuaiqi Huang , Liang Xiao , Shaochen Yang , and Xiang Li

**Abstract** To avoid the cases that control algorithm can't work well after application in green-house, and then cause losses, it's essential to test the performance of the control algorithm before putting it into practical use. To achieve this, we need a method to test the control algorithm rapidly and effectively. However, there is no such effective method to test the performance of the control algorithm in a greenhouse. With the rapid development of the IoT technology in the field of agricultural production, the method of process data modeling is proposed to help solve such problems. Based on the method of process data modeling, this paper proposes a new model which includes Z-transformation, scope control, OFF response and difference bias, and developed a greenhouse simulation test model. The experimental results show that our model can capture the changes of both temperature and humidity, and also limit the humidity within the set range. This proves that our model is effective and precise.

**Keywords** Z-transformation · Greenhouse simulation · Process data modeling

## 1 Introduction

Because the greenhouse environment simulation model plays an important role in the greenhouse structure design and the prediction of environmental changes [1], in order to realize the performance test of the greenhouse control algorithm in a short time, to avoid the losses caused by the poor performance of the algorithm after actual application. So it is of great significance to establish a greenhouse simulation test system based on the greenhouse environment simulation model to realize this function.

Greenhouse environment simulation models are generally structured based on mechanism models [1–3]. The mechanism model is based on a certain environmental factor in the greenhouse (such as temperature, etc.) or the mechanism change in the

---

Y. Jia · S. Huang · L. Xiao · S. Yang · X. Li ()

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

e-mail: [cqlixiang@cau.edu.cn](mailto:cqlixiang@cau.edu.cn)

greenhouse (such as the effect of crop transpiration rate, etc.) [1] to establish a mathematical model of the relevant dominant variables represented by the macro or micro equations [4], simulate the greenhouse environment. Since the greenhouse is a non-linear multi-input, multi-output complex system [5], its indoor environment will be affected by many indoor and outdoor factors. The response of the greenhouse system to external changes has different time scales and some factors are difficult to accurately measure and model [6, 7]. Therefore, it is very difficult to establish an accurate and effective greenhouse mechanism model. At the same time, the different physical parameters of the greenhouse will also vary depending on the facilities. Aging produces changes, causing decoupling control calculated parameters to deviate from the actual situation [6], and the model has poor adaptability to this change after the greenhouse environment changes [7], which makes it difficult to widely apply even if a good mechanism model is established.

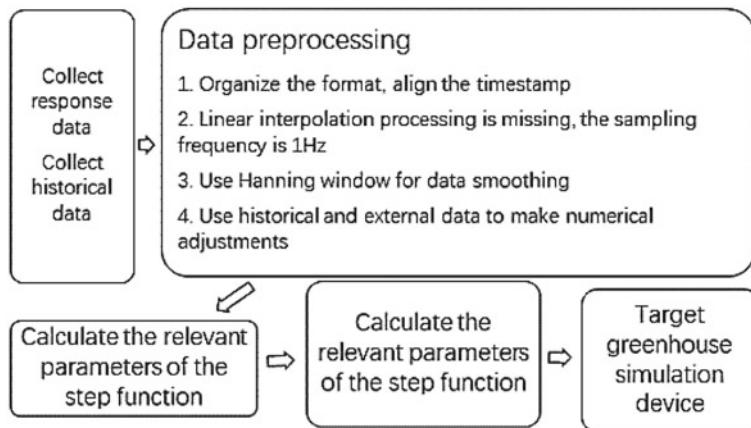
With the rapid development of the Internet of Things technology in the field of agricultural production, greenhouse data monitoring and signal control technologies have been relatively mature [8, 9], and the greenhouse control system's ability to obtain various aspects of information has rapidly improved. Therefore, a process model based on the operating data of the process can be considered to realize the real-time simulation of the process [10, 11]. This can reduce the difficulty of building test models and increase the speed of test model building while making full use of the rich data resources generated during the greenhouse operation.

This type of model is widely used in industry [12], and in the greenhouse, Peng et al. designed a greenhouse simulation model based on the process model based on the Z transformation [13]. However, in existing research, process modeling does not fully consider extreme situations. In actual production, extreme conditions, such as extreme temperatures, will cause irreversible and serious damage to agricultural products [14]. Since the data used in process modeling often comes from historical data [10–13], the amount of data in extreme cases is much smaller than the amount of data in normal cases, resulting in a lower degree of fitting of the model in extreme cases, increasing its probability of being misjudged and causing incorrect operation.

Therefore, this article uses limitate to control the range of the model based on the process model, and then uses the off response to avoid the severe vibration caused by the limitate, so as to realize that the model can still respond correctly under extreme conditions, and finally form a set Effective and simple greenhouse simulation test model.

## 2 Data Collection

The source of the data in this paper is the Sunlight Greenhouse of Zhuozhou Farm of China Agricultural University, through modeling the influence process of various operations in Zhuozhou Greenhouse. The steps of data collection and processing are shown in Fig. 1.



**Fig. 1** Greenhouse simulation data generation step diagram

The main controller of Zhuozhou Sunlight Greenhouse was operated, and the responses to various factors of the greenhouse under 10 different controls were collected. The specific response categories are shown in Table 1.

To ensure the time and data accuracy, special sensors are provided for data collection. Due to conditions, only the temperature and humidity data of the greenhouse were collected. The experiment carried out ten operations such as switching irrigation equipment, and the collection lasted for 270 min.

After obtaining the original data image, the data is preprocessed, and the steps are as follows:

- (1) Organize the data format and align the data timestamp;
- (2) Perform linear interpolation on the missing data segment and adjust the data sampling frequency to 1 Hz. The final data is valid for a total of 16,000 s;

**Table 1** Experimental control response category

Response category	Temperature/°C
Control category	Relative humidity/%
Pre-wait	
50% shade	
100% shade	
Turn off the sunshade	
Watering	
Watering off	
Down ventilation	
Turn off the ventilation	
Upper ventilation	
Close ventilation	

- (3) Use a Hanning window with a length of 150 to smooth the data, and remove the data whose length is the length of the time window before and after;
- (4) Adjust the collected greenhouse response data using historical data and the total characteristics of weather changes on the day, and try to keep only the target to control the influence of various factors in the greenhouse.

The original data change image of temperature and humidity and the preprocessed image are shown in Fig. 2. The left side is the original data curve, in which the blue line in the raw data graph represents the greenhouse data collected by the sensor, the green line represents the field sensor data near the greenhouse, the orange line represents the average data collected by the historical greenhouse sensor, and the gray line represents the time of the corresponding operation. The right side is the data change curve after preprocessing.

### 3 Model Design

It is known that the environmental factors in the greenhouse are affected by both the environmental changes in the greenhouse and the controller. Record the environmental change in the greenhouse at time  $t$  (the amount of change in temperature and humidity per unit time) as  $\Delta(t)$ , and the simulated value of the changes in greenhouse factors after the greenhouse control changes at time  $t$  is  $\Phi(t)$ , the adjusted proportional coefficient is  $k$ , the simulated value of temperature and humidity at time  $t$  is  $y_i(t)$ . The specific simulation formula is:

$$y_i(t) = \Phi(t) + k\Delta(t) \quad (1)$$

$\Phi(t)$  can be derived from the model function. The form of function  $G$  is as follows:

$$\begin{aligned} \begin{bmatrix} y_{tem} \\ y_{ri} \end{bmatrix} &= G \cdot [u_1 \ u_2 \ \dots \ u_n]^T, \\ G &= \begin{bmatrix} G_{11} \ \dots \ G_{1n} \\ G_{21} \ \dots \ G_{2n} \end{bmatrix}, \\ G_{ij}(s) &= \frac{k_{ij}}{T_{ij}s+1} e^{-\tau_j}, \quad n = 10 \end{aligned} \quad (2)$$

In the formula (left),  $\sim$  represents the input of 10 control signals (as shown in Table 1). And respectively represent the temperature response and humidity response after adjustment.  $G$  is the transfer function matrix, and each item in the matrix corresponds to the transfer function of a certain input to a certain response. The independent variable of the transfer function is time, and the dependent variable is the ratio of the Laplace transform of the output waveform to the Laplace transform of the input waveform.

The transfer function form (the above formula (right)) contains three parameters  $k$ ,  $T$ , and  $\tau$ , which correspond to the change amplitude, change time and lag time of

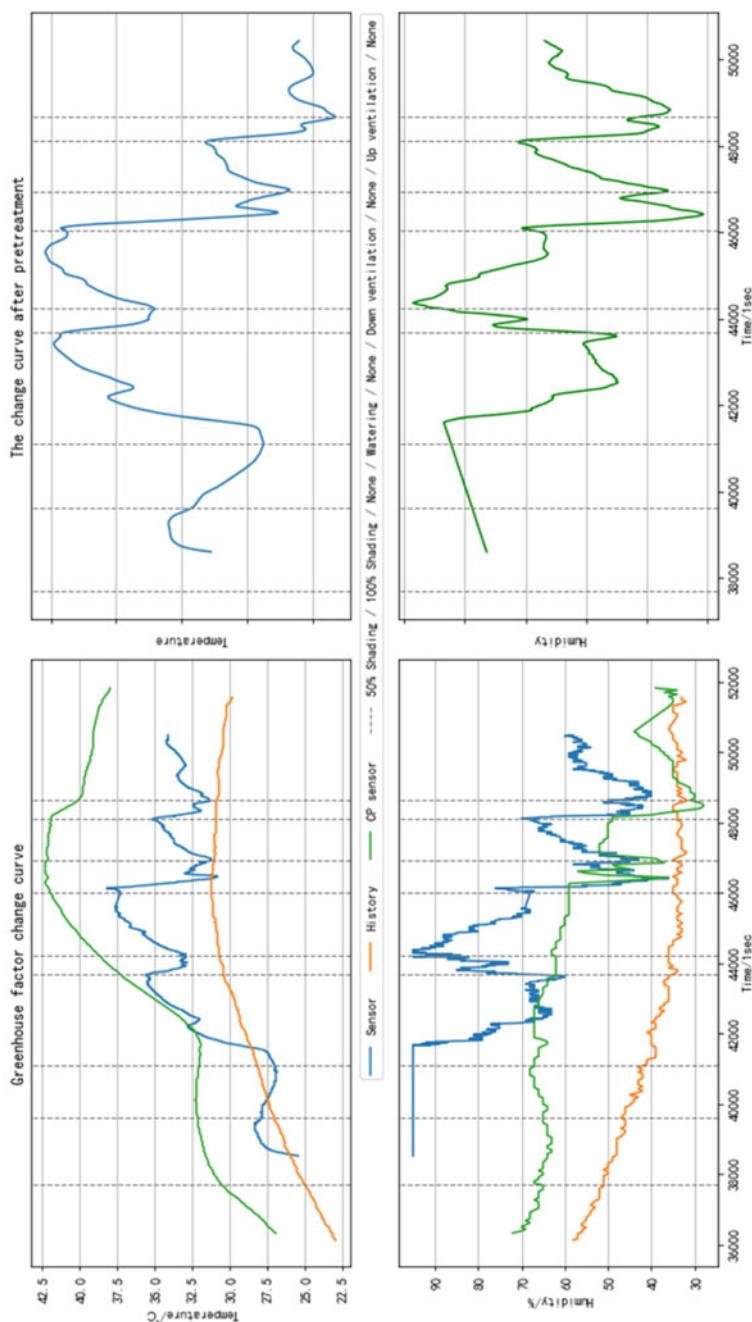


Fig. 2 The change curve of various factors in the greenhouse before and after preprocessing

each control response in the actual scene. Through the collected temperature, room temperature and humidity curve, using the calculation method provided by Peng et al. [12], the values of the three parameters of each transfer function can be obtained. The calculation formula is as follows, and the calculation results are shown in Table 2.

$$\begin{cases} T = \frac{t_2 - t_1}{\ln(1 - \hat{y}(t_1)) - \ln(1 - \hat{y}(t_2))} \\ \tau = \frac{t_2 \ln(1 - \hat{y}(t_1)) - t_1 \ln(1 - \hat{y}(t_2))}{\ln(1 - \hat{y}(t_1)) - \ln(1 - \hat{y}(t_2))} \end{cases}, \quad \hat{y}(t) = \frac{y(t)}{y(\infty)} \quad (3)$$

Since the above formula is a continuous function in time sequence, and the sensor sampling data is discrete data, the formula needs to be discretized. Discretization of the z-transformation of this formula is as follows:

$$y_i(z) \prod_{j=1}^n \left( z - e^{-\frac{T}{T_{ij}}} \right) = \sum_{j=1}^n \left[ \frac{k_{ij}}{T_{ij}} \cdot \prod_{k=1, k \neq j}^n \left( z - e^{-\frac{T}{T_{ik}}} \right) \cdot z^{1 - \frac{\tau_{ij}}{T}} \cdot u_j(z) \right] \quad (4)$$

Perform z inverse transformation and sort it out as follows (this time the greenhouse sensor acquisition frequency is 200 s/time, so take):

**Table 2** Response parameter calculation result

Control category	Temperature response parameter			Humidity response parameters		
	k	T	$\tau$	k	T	$\tau$
Pre-wait	2.9510572	779.61752	288.63793	-1.835297	700.46854	271.76098
50% shade	0.7059091	39.574494	952.43847	2.2853996	463.02157	1078.1301
100% shade	-2.306273	542.17056	242.00709	3.531767	718.27706	208.95829
Turn off the sunshade	6.1859365	397.72366	448.40666	-24.87508	484.78755	498.3713
Watering	-2.822252	102.89368	68.140031	24.919088	104.87241	27.161955
Watering off	2.6401004	352.21299	79.902413	-15.18066	263.17038	749.91585
Down ventilation	-6.499281	132.57455	190.46889	-17.48419	49.468117	166.54809
Turn off the ventilation	2.2648786	340.34064	115.77087	21.518854	508.53224	154.63438
Upper ventilation	-3.850253	114.76603	70.271573	-18.55061	57.383016	60.635786
Close ventilation	1.0049377	104.87241	214.16195	14.23492	148.40435	723.64428

$$\begin{aligned}
y_i(k) = & \sum_{j=1}^n \left\{ \frac{k_{ij}}{T_{ij}} \cdot \left[ u_j \left( k - \frac{\tau_{ij}}{T} \right) + u_j \left( k - 1 - \frac{\tau_{ij}}{T} \right) \cdot \right. \right. \\
& \left. \left. \sum_{k=1, j \neq k}^n -e^{-\frac{T}{T_{ij}}} + \dots + u_j \left( k - n + 1 - \frac{\tau_{ij}}{T} \right) \cdot e^{k=\sum_{j=1, j \neq k}^n -\frac{T}{T_{ik}}} \right] \right\} \\
& - \left[ y_i(k-1) \cdot \sum_{j=1}^n -e^{-\frac{T}{T_{ij}}} + y_i(k-2) \cdot \sum_{j=1, k=1, j \neq k}^n e^{-\frac{T}{T_{iv}} - \frac{T}{T_{it}}} + \dots + y_i(k-n) \cdot e^{i=1 - \frac{T}{T_{ij}}} \right]
\end{aligned} \tag{5}$$

Because the relative humidity response has a value range, the response variable is added to the transfer function with a value range response. Assuming that a certain response range is, the corresponding transfer function has the following form:

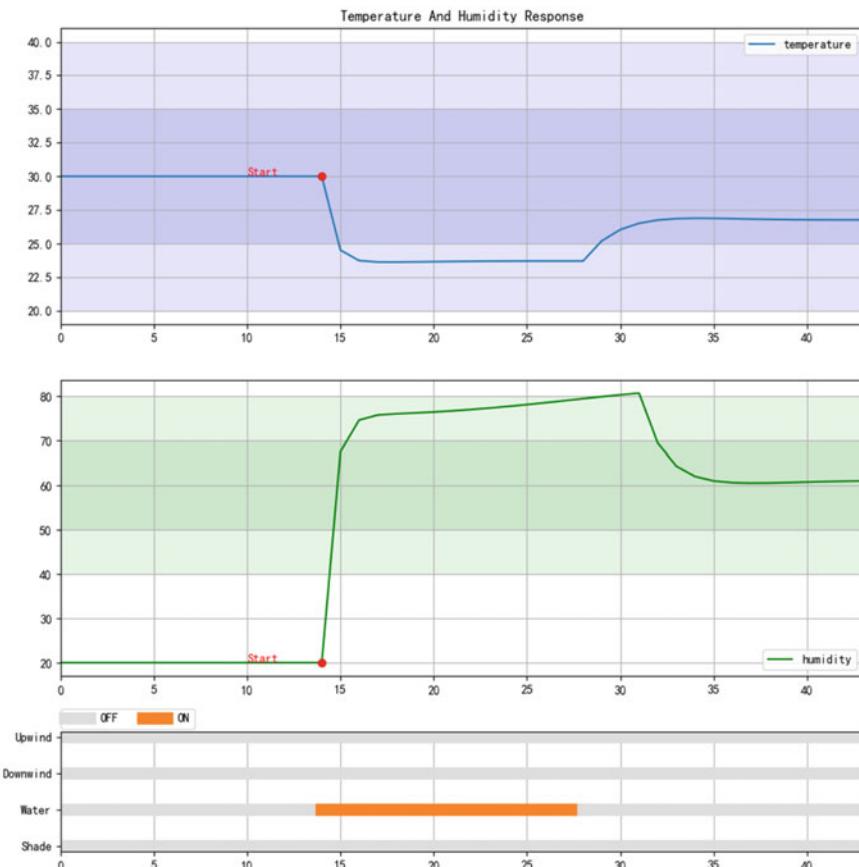
$$\begin{aligned}
y_i(k) = & \text{limitate} \left( \sum_{j=1}^n \left\{ \text{isOpen}_i \otimes \text{isChange}_{ij} \cdot \frac{k_{ij}}{T_{ij}} \cdot \left[ u_j \left( k - \frac{\tau_{ij}}{T} \right) + u_j \left( k - 1 - \frac{\tau_{ij}}{T} \right) \cdot \right. \right. \right. \\
& \left. \left. \left. \sum_{k=1, j \neq k}^n -e^{-\frac{T}{T_{ij}}} + \dots + u_j \left( k - n + 1 - \frac{\tau_{ij}}{T} \right) \cdot e^{k=\sum_{j=1, j \neq k}^n -\frac{T}{T_k}} \right] \right\} - (1 - \text{isOpen}_i) \cdot y_i(k-1) \right. \\
& - \text{isOpen}_i \cdot \left[ y_i(k-1) \cdot \sum_{j=1}^n -e^{-\frac{T}{T_{vij}}} + y_i(k-2) \cdot \right. \\
& \left. \left. \sum_{j=1, k=1, j \neq k}^n e^{-\frac{T}{T_{ij}} - \frac{T}{T_{ik}}} + \dots + y_i(k-n) \cdot e^{i=1 - \frac{T}{T_{ij}}} \right] \right) \\
& \text{limitate} = x : \max(a, \min(x, b)) \\
& \text{is Open}_i = y_i(k-1) \otimes y_i(k-2) \\
& \text{is Change}_{ij} = u_j \left( k - \frac{\tau_{ij}}{T} \right) \otimes u_j \left( k - 1 - \frac{\tau_{ij}}{T} \right) \\
& \otimes = x, y : \begin{cases} 1 & x \neq y \\ 0 & x = y \end{cases}
\end{aligned} \tag{6}$$

The formula  $y(k)$  is the y response value at the  $k$ th time, which can be solved by the y response value and the control before the  $k$  time; the function  $u_j(x)$  refers to the value of the  $j$ th control at the  $x$ th time; the value corresponding to each control will be 0 At the beginning, when the control is carried out, the corresponding value will be accumulated by 1, so as to realize the superimposition of the control; limitate is used to limit the value of the existing factors in the greenhouse, such as limiting the value of relative humidity between 0 to 100; IsOpen and IsClose are used to close the response and avoid the problem of control failure caused by severe oscillation caused by limitate.

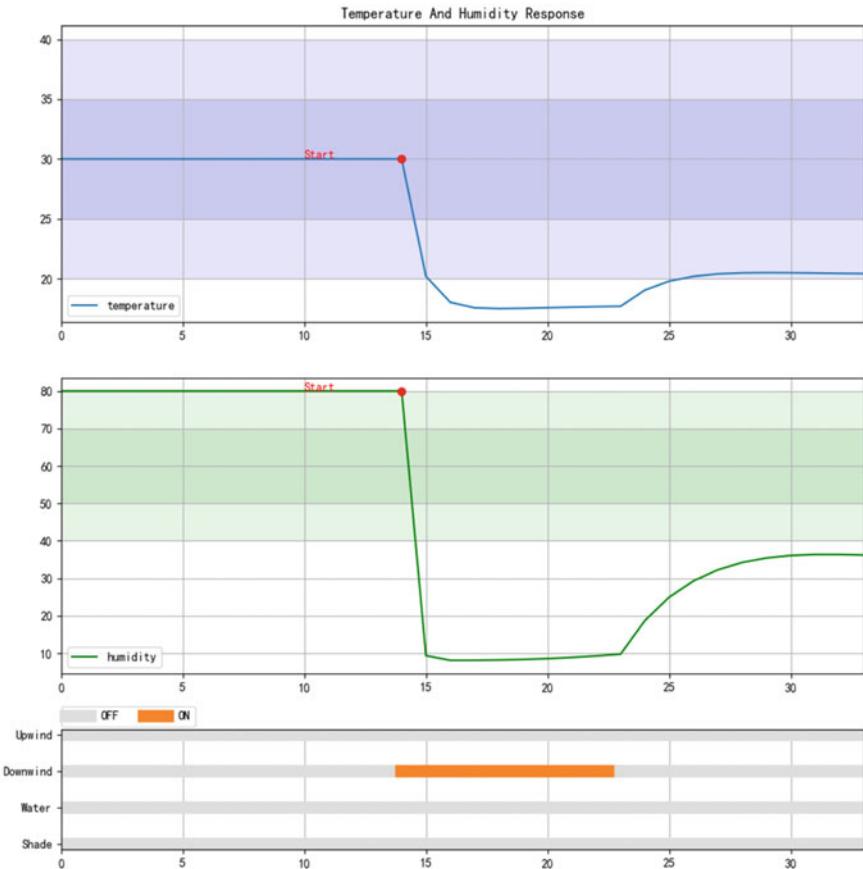
## 4 Simulation

In order to verify whether the model change is consistent with the actual situation, a response experiment is first performed on a single control device. The specific process is as follows:

- (1) Experiment with on-off watering. As shown on the left side of Fig. 3, it is divided into three parts from top to bottom, namely the temperature change curve, the humidity change curve and the corresponding controller status. When watering at the 13th moment, the humidity rises and the temperature drops; when the watering stops at the 27th moment, the humidity drops and the temperature rises. The experimental results are consistent with the effect of watering on temperature and humidity.



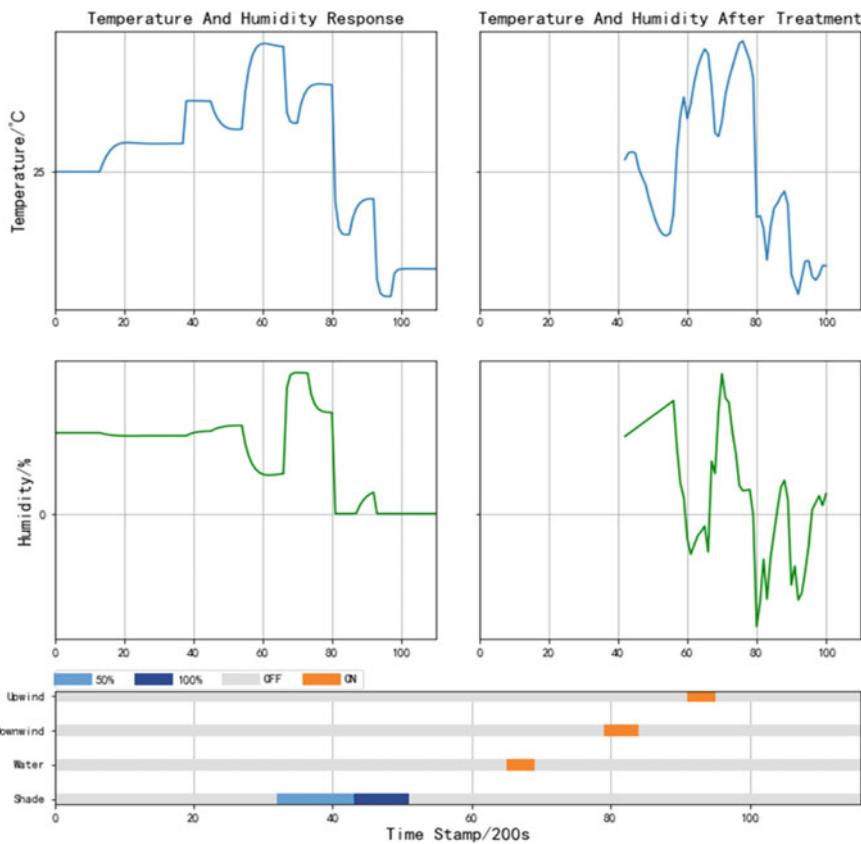
**Fig. 3** Single controller simulation response experiment



**Fig. 3** (continued)

- (2) Experiment with ventilation under the switch. As shown on the right side of Fig. 3, when the lower ventilation window is opened at the 13th moment, the humidity drops and the temperature drops; when the ventilation stops at the 23rd moment, the temperature and humidity begin to rise. The experimental results are consistent with the effect of ventilation on temperature and humidity.

The final simulation result of the original greenhouse scene is shown in Fig. 4, where the left side is the simulation result, and the right side is the real data. It can be found that the variation trend of each response of the simulation results is roughly the same as that of the greenhouse temperature and humidity after pretreatment, and after the limitate range control of the simulation model, the simulated value of humidity does not fall below the actual 0 at the 80th time. But kept in the range of 0–100 set above, so the data obtained by simulation can be used for subsequent control verification.



**Fig. 4** Simulation results of the original greenhouse scene

In summary, the constructed greenhouse temperature and humidity change model is effective and accurate.

## 5 Conclusion

Based on the process modeling method, this paper constructs a set of greenhouse simulation test models that can simulate the environmental changes after control through Z transformation, limitate range control, off response and differential bias. It can be seen from the results of the simulation experiment that the simulation of this model is effective and accurate.

## References

1. Weimin, D., Xiaoxiong, W., Yinian, L., Jian, W.: Analysis of research status of greenhouse environment control and greenhouse simulation model[J]. Transactions of the Chinese Society of Agricultural Machinery **40**(05), 162–168 (2019)
2. Lihong, X., Yuaping, S., Yuming, L.: Requirements and status quo of control-oriented greenhouse system microclimate environment model[J]. Transactions of the Chinese Society of Agricultural Engineering **29**(19), 1–15 (2013)
3. Lili, M., Qichang, Y., Gerard, B., Nan, W.: Construction of simulation model of thermal environment in solar greenhouse[J]. Transactions of the Chinese Society of Agricultural Engineering **25**(01), 164–170 (2009)
4. Pengfei, C., Xionglin, L.: Research progress of soft sensor modeling methods in chemical process[J]. CIESC Journal **64**(03), 788–800 (2013)
5. Dan, L., Xin, C., Chongwei, H., Liangliang, J.: Intelligent Agriculture Greenhouse Environment Monitoring System Based on IOT Technology (2015)
6. Lujuan, D., Kanyu, Z., Youmin, G., Chunhong, C.: Preliminary study on multi-level control system and optimized target value setting of greenhouse environment[J]. Trans. Chinese Soc. Agricult. Eng., pp. 119–122 (2005)
7. Xuehua, Z., Wu, Z., Xu, Y., Lujiao, W., Huimin, M., Qiong, F.: A review of research on environmental control methods of agricultural greenhouses[J]. Control. Eng. **24**(01), 8–15 (2017)
8. Li, P., Wang, J.: Research progress of intelligent management for greenhouse environment information[J]. Nongye Jixie Xuebao/Transactions of the Chinese Society for Agricultural Machinery **45**, 236–243 (2014)
9. Qin, L., Lu, L., Shi, C., Wu, G., Wang, Y.: Implementation of IOT-based greenhouse intelligent monitoring system[J]. Nongye Jixie Xuebao/Trans. Chinese Soc. Agricult. Mach. **46**, 261–267 (2015)
10. Chuanhou, G., Ling, J., Jiming, C., Youxian, S.: Data-driven modeling and prediction algorithm for complex blast furnace ironmaking process[J]. Acta Automatica Sinica **35**(06), 725–730 (2009)
11. You, L.: Research and application of modeling method based on process data[D]. North China Electric Power University (2014)
12. Min, L.: Summary of research on data-based production process scheduling methods[J]. Acta Automatica Sinica **35**(06), 785–806 (2009)
13. Xuan, P.: Research and development of decoupling control system for greenhouse environment variables based on neural network[D]. Xinjiang University (2018)
14. Rachel, F., Ellis, R. H., Wheeler, T. R., et al.: Effect of High Temperature Stress at Anthesis on Grain Yield and Biomass of Field-grown Crops of Wheat[J]. Annals of Botany, pp. 631–639 (1998)

# Vis–NIR Hyperspectral Dimensionality Reduction for Nondestructive Identification of China Northeast Rice



Jiahao Wang , Chun Liao , Jingyi Zhao , and Wanlin Gao

**Abstract** The establishment of a nondestructive identification model for China Northeast Rice is of great importance for market consumption. nondestructive identification of 3 types of China Northeast Rice using Vis–NIR hyperspectral images. A visible near-infrared hyperspectral system (382.19 ~ 1026.66 nm) was used to collect 900 rice data, and the region of interest (ROI) was determined by ENVI, and the ROI regions of three rice were treated as one sample, and finally, 300 sample data were obtained. First, the sample data is preprocessed using standard normal variable transformation (SNV). Then competitive adaptive reweighted sampling (CARS), successive projection algorithm (SPA), principal component analysis (PCA), and partial least squares (PLS) are used to reduce the dimensionality of the preprocessed data. Finally, the reduced-dimensional features are fed into the linear discriminant analysis (LDA) for training. The results show that the data can reduce the complexity of the model after dimensionality reduction. The model based on SVN, PLS and LDA has the best accuracy of 91.67% on the Validation, it can effectively replace the full wavelength data for nondestructive identification of China Northeast Rice.

**Keywords** Rice · Vis–NIR · Hyperspectral · Dimensionality reduction

## 1 Introduction

China is one of the largest rice producers and rice exporters in the world, and rice is also one of the most common major food crops in China [1]. Rice is rich in carbohydrates and can also provide the body with nutrients such as protein, fat and vitamins. With the continuous improvement of people's living standards in our country, they pay more attention to the "taste", color and quality of rice, but there are

---

J. Wang · C. Liao · J. Zhao · W. Gao ()

Key Laboratory of Agricultural Information Standardization, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China  
e-mail: [gaowl@cau.edu.cn](mailto:gaowl@cau.edu.cn)

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

huge differences in the price of different types of rice. Due to the imperfect production standards of my country's agricultural product market, many unscrupulous traders use shoddy rice in the process of selling rice in order to obtain more benefits, which seriously damages the interests of consumers and leads to chaos in the rice market [2].

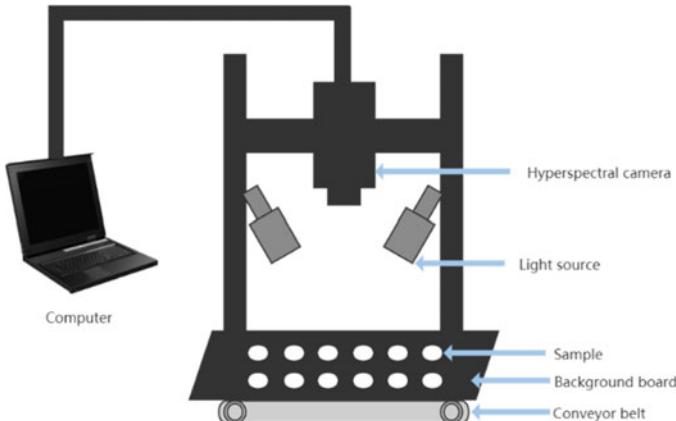
The traditional methods of identifying rice include sensory identification, stable isotope technique identification, mineral element analysis technique identification, biological identification and intelligent sensory biomimetics identification, etc. but these methods are subjective, destructive and long-period. Cannot perform batch appraisal [3, 4]. IQBAL [5] et al. studied the differences in vitamin E and aflatoxin content in different rice varieties. Qian [6] et al. studied the differences in thiamin and riboflavin contents in different rice varieties. Li Wenbing [7] et al. established inductively coupled plasma mass spectrometry to determine the content of various mineral elements in rice from different origins, and found that there were significant differences in various mineral elements in different rice. The above research provides a basis for the identification of rice using NIR spectroscopy technology. However, due to the small size of the rice and the uneven distribution of the ingredients, NIR spectroscopy generally requires pulverizing the sample into powder, and it is impossible to collect the spectral data in a nondestructive manner. Vis–NIR hyperspectral images combine the spectrum and images of the visible wavelength, contains very rich spectrum and spatial information, and can collect the spectrum data of the sample non-destructively.

Using Vis–NIR hyperspectral imaging technology, three kinds of famous and high-quality rice were selected for nondestructive identification, which provided experimental basis for establishing a nondestructive rice identification model. The hyperspectral images of rice from 382.19~1026.66 nm were collected, and the spectral data were preprocessed by SNV. In order to build a more accurate model, CARS, SPA, PCA and PLS are used to reduce the dimensionality of the preprocessed data to verify the performance of LDA. At the same time, a Vis–NIR hyperspectral method is proposed for nondestructive identification of rice in current application scenarios.

## 2 Experimental Section

### 2.1 Sample Preparation

This paper used three kinds of famous and high-quality China Northeast Rice, including Meihe rice (Meihekou City, Jilin Province), Wuchang Rice (Wuchang City, Heilongjiang Province), and Panjin Rice (Panjin City, Liaoning Province). Three kinds of rice were purchased from the place of production, and 300 were randomly selected from each of the purchased rice, for a total of 900.



**Fig. 1** Schematic of the Vis-NIR hyperspectral imaging system

## 2.2 Vis-NIR Hyperspectral Imaging System

The GaiaSorter hyperspectral sorter was used for the data acquisition, and Image- $\lambda$ -V10E-LU of ZOLIX INSTRUMENTS CO.,LTD was used for camera. The spectral range is 382.19~1026.66 nm, the wavelength spacing is 0.84 nm, and a total of 728 wavelengths. The parameter settings for data acquisition include the distance between the lens of the camera and the rice is 18 cm; the platform movement speed is 0.5 cm/s; the integration time is 9.6 ms. The schematic diagram of the Vis-NIR hyperspectral imaging system is shown in Fig. 1.

## 2.3 Image Acquisition and Correction

Before the measurement, the instrument was turned on for 30 min to stabilize the light. As shown in Fig. 1, the non-overlapping and uniformly spaced samples were placed on a mobile platform, and the Vis-NIR hyperspectral equipment was used to collect data.

To reduce the uneven distribution of dark current and light source intensity caused by the long-term use of the camera, it is necessary to perform black and white correction on the obtained Vis-NIR hyperspectral image. The calculation method is Eq. (1).

$$I = \frac{I_{raw} - I_{Dark}}{I_{White} - I_{Dark}} \quad (1)$$

In Eq. (1),  $I$  is the corrected image;  $I_{raw}$  is the original image;  $I_{Dark}$  is the blackboard image;  $I_{White}$  is the whiteboard image.

## 2.4 Spectra Data Exaction

In this paper, the software ENVI 5.3 was used to select the region of interest (ROI) for 900 rice. Due to the influence of noise and the small rice volume during the data collection process, the whole rice is regarded as one ROI, and the 3 ROI regions of each type of rice are regarded as one sample. The average spectral data is obtained by calculating the average reflectivity of the pixels in the sample ROI. Repeat the same steps for 3 kinds of samples, and finally obtain a 300\*728 spectral matrix of 300 samples. The 300 samples were randomly divided into 4:1 training set and test set, of which 240 samples were in the training set and 60 samples were in the test set.

## 2.5 Preprocess

SNV can reduce the influence of factors such as light scattering, baseline shift and low signal-to-noise ratio of the system on the data when collecting hyperspectral images of the sample [8]. The calculation method is Eq. (2).

$$X_{SNV} = \frac{X - \mu}{\sigma} \quad (2)$$

In Eq. (2),  $X_{SNV}$  is the spectral data after SNV,  $X$  is the original spectrum of the sample,  $\mu$  is the mean value, and  $\sigma$  is the standard deviation.

## 2.6 Feature Selection

**Competitive adaptive reweighted sampling (CARS).** The principle of CARS is to mimic the Darwinian evolutionary theory of “survival of the fittest”. It can remove variables with no reference value while minimizing the influence of covariates on the model. The algorithm is as follows:

- (1) Monte Carlo sampling (MCS): a certain proportion of sample data is randomly selected from the sample set to build a PLS model, and the weight ( $W_i$ ) of the  $i$ th variable is calculated by Eq. (3).

$$W_i = \frac{|\beta_i|}{\sum_{i=1}^p |\beta_i|}, i = 1, 2, \dots, p \quad (3)$$

In Eq. (3),  $\beta_i$  is the regression coefficient of the PLS model, and  $p$  is the number of wavelengths in the original sample set.

- (2) Screening weight  $W_i$  wavelength values, based on the exponential decay function to eliminate  $|\beta_i|$  smaller wavelengths, retain  $|\beta_i|$  larger wavelengths, the retention rate is calculated by the exponential function Eq. (4)

$$r_i = ae^{-ki} \quad (4)$$

In Eq. (4),  $a = \left(\frac{P}{2}\right)^{\frac{1}{N-1}}$ ,  $k = \frac{[\ln(\frac{P}{2})]}{N-1}$ ,  $N$  is the number of MCS.

- (3) A subset of  $N$  wavelengths is obtained after  $N$  MCSs, and the subset of variables with the smallest RMSECV in each MCS process is taken as the optimal subset of wavelengths.

**Successive projection algorithm (SPA).** SPA uses projection analysis of vectors to find the variable group with the least redundant information in the original hyperspectral data, so as to minimize the correlation between the variables in the group. It selects the variable combination that summarizes most of the sample information and the variable with the smallest covariance. The algorithm is as follows:

- (1) Randomly select the  $j$ th column of the original spectrum matrix and assign it to  $x_j$ , denoted as  $x_{k(0)}$ .
- (2) The set of remaining column vectors in the spectral matrix is denoted as  $S$ ,  $S = \{J, 1 \leq j \leq J, j \notin \{K(0), \dots, K(n-1)\}\}$ .
- (3) Calculate the projection of  $x_j$  on the remaining column vector according to the Eq. (5):

$$P_{xj} = x_j - (x_j^T x_{k(n-1)}) x_{k(n-1)} (x_{k(n-1)}^T x_{k(n-1)})^{-1}, j \in s \quad (5)$$

- (4) Select the spectral wavelength of the largest projection vector in step 3):

$$K(n) = \arg(\max(|P(x_j)|, j \in s)) \quad (6)$$

- (5) Let  $x_j = p_x, j \in s$ .  
 (6) Let  $n = n + 1$ , if  $n < N$ , then go back to step 2) to cycle the calculation.

In the above steps,  $x_{k(0)}$  is the original iteration vector,  $N$  is the number of wavelengths selected, and  $J$  is the number of columns of the original spectral matrix. The final wavelengths combination obtained by selection is  $\{x_{k(n)} = 0, \dots, N-1\}$ . The  $k(0)$  and  $N$  corresponding to the RMSECV obtained through multiple linear regression analysis in each cycle are the optimal values.

## 2.7 Feature Extraction

**Principal component analysis (PCA).** PCA is an unsupervised linear algorithm that can simultaneously solve the problem of effective information retention and multicollinearity between variables. The principle is to map the original  $n$ -dimensional

variables to k-dimensions by a linear transformation, and the information of these k-dimensional new variables do not overlap and are orthogonal to each other. It has been widely used in dimensionality reduction and decorrelation of hyperspectral data.

**Partial least squares (PLS).** PLS is a supervised algorithm that combines the advantages of linear regression, canonical correlation analysis and PCA. It can reduce the information in the original data that is irrelevant to the predicted values while maximizing the correlation between the hidden information in the original data and the predicted values.

## 2.8 Feature Extraction

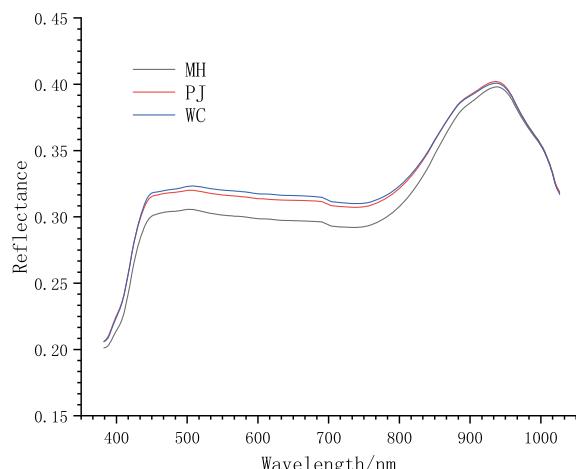
This paper used LDA applied in various research fields to build the model. The idea is: after projecting the data, the variance between different categories is the largest, and the variance between the same categories is the smallest, and then the threshold is used for discriminative classification.

## 3 Results and Discussion

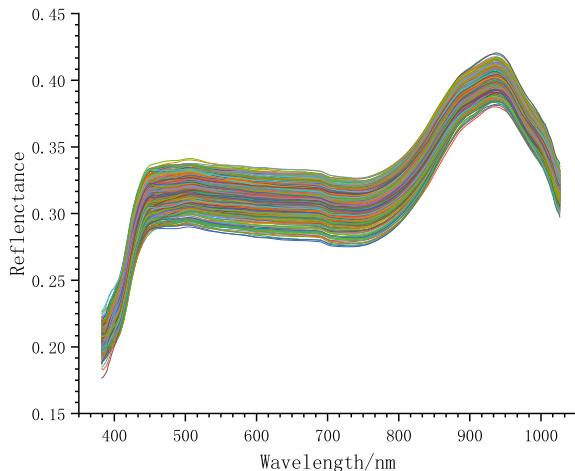
### 3.1 Spectral Features of Rice

In this paper, the average reflectance of the three types of rice used at all wavelengths (382–1027 nm) is shown in Fig. 2. It can be seen that the reflectance corresponding

**Fig. 2** MH: Meihe rice, PJ: Panjin rice, WC: Wuchang rice



**Fig. 3** The original reflectance of different rice



to each wavelength of different rice is different, but the average reflectance curve has similar trends, which means that the percentages of various components are different. Factors such as planting environment, cultivation process and temperature difference between day and night may cause these differences.

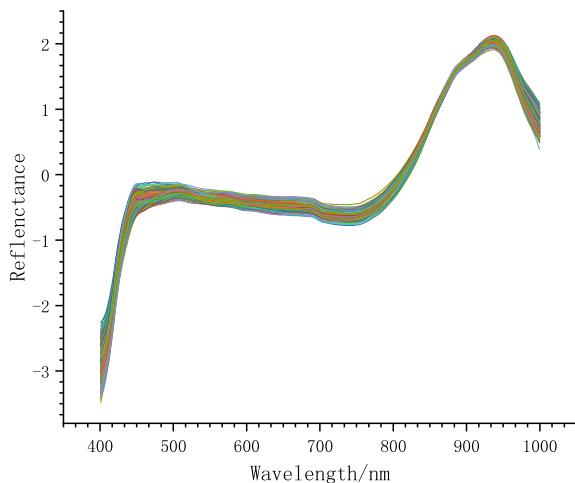
### 3.2 The Results of Preprocessing

In the hyperspectral image data acquisition stage, the influence of some external factors had been reduced by black-and-white correction, but there are still various external factors in the obtained spectral data, and these noises will cause a certain gap between the predicted value and the real value. From the original reflectance in Fig. 3, it can be seen that the data is greatly disturbed at the beginning and end of the collection. Therefore, this paper excluded the first 22 wavelengths and the end 29 wavelengths, and the final wavelength range used was 400.78~999.77 nm. The remaining wavelengths were preprocessed using SNV, and the processed reflectance are shown in Fig. 4.

### 3.3 The Results of Dimensionality Reduction

The preprocessed data is processed by dimensionality reduction. CARS sets the number of MCS to 50 times. MCS sampling has random. Therefore, it is necessary to compare the RMSECV values through repeated iterations. When the minimum RMSECV value is 0.412, the number of characteristic wavelengths at this time is 44. When SPA is used for feature selection, the minimum number of feature wavelengths

**Fig. 4** The reflectance of rice after SNV



is set to 20, and when the minimum RMSECV value is 0.407, 20 feature wavelengths are obtained. PCA reduced the input features to 9 principal components, and the variance contribution rate of the first 9 principal components was 99.47%. Using PLS to analyze the full wavelength of rice, the original data were reduced to 12 features.

It can be seen from Table 1 that different dimensionality reduction results can be obtained under different dimensionality reduction algorithms, and the original 677-dimensional data can be reduced to dozens of dimensions after dimensionality reduction.

**Table 1** Results of dimensionality reduction by different algorithms

Algorithm	Number of variables	Characteristic wavelength (nm)
CARS	44	410.10, 410.95, 413.49, 423.68, 427.08, 438.14, 455.20, 466.31, 467.17, 468.03, 479.16, 503.21, 504.07, 548.99, 549.85, 552.45, 595.96, 596.84, 597.71, 599.46, 600.33, 674.13, 675.01, 675.90, 690.94, 691.82, 692.71, 693.60, 724.68, 763.06, 800.77, 802.57, 803.47, 813.38, 815.18, 881.34, 899.57, 924.26, 937.10, 949.95, 950.87, 960.07, 984.97, 999.77
SPA	20	423.68, 444.96, 555.06, 598.58, 651.20, 772.92, 815.18, 846.82, 872.24, 889.54, 900.48, 903.22, 905.05, 907.79, 912.36, 935.26, 944.44, 951.79, 960.07, 999.77
PCA	9	\
PLS	12	\

**Table 2** Results of different models

Experiment	Method	Number of variables	Calibration (%)	Validation (%)	modeling and testing time (s)
Model 1	SNV + LDA	677	100.00	75.00	0.174
Model 2	SNV + CARS + LDA	44	98.33	88.33	0.008
Model 3	SNV + SPA + LDA	20	92.50	85.00	0.003
Model 4	SNV + PCA + LDA	9	81.25	80.00	0.001
Model 5	SNV + PLS + LDA	12	93.75	91.67	0.002

### 3.4 Comparison of Models

By comparing the experiments, the results of different models were obtained as shown in Table 2. From the results of model 1 (75.00%), the full wavelength LDA classification algorithm based on SNV preprocessing can distinguish different China Northeast Rice to a certain extent.

From the results of Model 1 (75.00%), Model 2 (88.33%), Model 3 (85.00%), Model 4 (80.00%), and Model 5 (91.67%), the models using the dimensionality reduction method all improved on the accuracy of Model 1. The 677 features of the original data could be reduced to dozens of features, which can also indicate that there was some redundancy in the original data. At the same time, the modeling and testing time was relatively short.

The results of Model 2 (88.33%), Model 3 (85.00%), Model 4 (80.00%) and Model 5 (91.67%) show that the model has the highest accuracy among the dimensionality reduction methods using PLS. And the model after using PLS was also relatively superior from the perspective of model complexity. The results indicate that the nondestructive identification model of China Northeast Rice established by PLS dimensionality reduction is feasible.

The results of using LDA could also reflect the strong linearity of the data set in this paper, and the good performance of the results of using LDA could be expected.

## 4 Conclusions

It could be seen that the Vis–NIR hyperspectral images can realize the nondestructive collection of sample data. Through theoretical analysis combined with experimental verification, and by comparing different dimensionality reduction methods, an effective method for processing hyperspectral data was proposed. First, use SNV preprocessing to reduce the impact of the sampling process, and then use CARS, SPA,

PCA, and PLS to reduce the complexity of the model. Finally, the input features are classified by the LDA. Among them, the models based on SNV, PLS and LDA had achieved the best performance in the current application scenario of China Northeast Rice nondestructive identification, with an accuracy rate of 91.67% on the Validation. The results show that the models established by SNV, PLS and LDA are effective for nondestructive identification of China Northeast Rice based on hyperspectral, and the use of PLS for dimensionality reduction instead of the full wavelengths could effectively perform the nondestructive identification of China Northeast Rice.

## References

1. Lu, W., Dan, L., Hongbin, P., et al.: Use of hyperspectral imaging to discriminate the variety and quality of rice. *Food Anal. Methods* **8**(2), 515–523 (2015)
2. Long, L., Jingzhu, W., Cuiling, L., et al.: Research on the hyperspectral identification method of rice producing area in northeast/african northeast based on model cluster. *Spectroscopy Spectral Anal.* **40**(3), 905–910 (2020)
3. Shengying, H., Hongbo, R., Jun, Z., et al.: Research progress of rice producing area traceability method. *Chin. Agric. Sci. Bull.* **36**(14), 148–155 (2020)
4. Jing, W., Zhenyu, Y., Yao, Z., et al.: Research on identification methods and standardization of Wuchang rice. *Agricult. Products Process.* **17**, 46–49 (2016)
5. Iqbal, S.Z., Mustafa, H.G., Asi, M.R., et al.: Variation in vitamin E level and aflatoxins contamination in different rice varieties. *J. Cereal Sci.* **60**(2), 352–355 (2014)
6. Yongwen, Q., Junzan, L., Kunming, H., et al.: Vitamin B1 and B2 content in different varieties of rice. *Acta Agron. Sin.* **17**(1), 58–63 (1991)
7. Wenbing, L., Qi, Z., Zhendu, Z., et al.: Inductively coupled plasma mass spectrometry analysis of multiple mineral elements in rice from different origins. *Northeast Agricult. Sci.* **45**(06), 129–134 (2020)
8. Rinnan, A., Van Den Berg, F., Engelsen, S.B.: Review of the most common pre-processing techniques for near-infrared spectra. *Trac-Trends Anal. Chem.* **28**(10), 1201–1222 (2009)

# An Encryption Scheme for Internet of Things Monitoring System



Haoyi Sun , Shuihai Zhang , Chunli Lv , and Bei Pei

This work was supported by the Key Laboratory of Information and Network Security, Ministry of Public Security, the Third Research Institute of the Ministry of Public Security(C19605).

**Abstract** Aiming at the security problem of the Internet of Things monitoring system, based on two algorithms of blind decryption and threshold RSA signature, this paper proposes an encryption scheme for the Internet of Things system for sensors such as cameras. This solution double-encrypts the recorded video information before the camera sends it to the cloud service provider for storage, and shares the private key for decryption at the key escrow center in a secret sharing manner. This solution can not only effectively protect the confidentiality and integrity of video information, but can also be used to ensure information security for other IoT systems. This article also analyzes the security of the encryption scheme from different perspectives. This scheme is a relatively novel encryption scheme for the Internet of Things system, which can meet the needs of users to store and access information through the Internet of Things system in the current era.

**Keywords** Internet of things · Blind decryption · Threshold signature · Monitoring system

## 1 Introduction

In the past, the traditional video surveillance system requires the central system platform, video-bearing network, front-end camera as well as storage devices, such series of hardware facilities has greatly increased the construction cost. Now, relying on the Internet of things technology, the surveillance video can be uploaded to the cloud, which means it merely requires customers to buy cameras to achieve the monitoring

---

H. Sun · S. Zhang · C. Lv ()

College of Information and Electrical Engineering, China Agricultural University, Beijing, China  
e-mail: [lvcl@cau.edu.cn](mailto:lvcl@cau.edu.cn)

B. Pei

The Third Research Institute of the Ministry of Public Security, Shanghai, China

function like that of the traditional monitoring system after some simple settings. The now prevalent Internet of things monitoring system is mainly implemented by the Internet enterprises which provide system service platform. Users only need to buy smart cameras, then register their accounts and set up their own cameras on the system service platform provided by cloud service providers. Thereby, the camera will be able to upload the information of the collected videos to the cloud. When users want to access and view the monitoring information of the camera, they simply need to log in to the system service platform of the cloud service provider through any smart terminal such as mobile phone, tablet computer, and so on, so that they can view the field information of their concern anytime and anywhere.

However, this kind of surveillance system is different from the video surveillance system in public places. The user's video information is stored in the cloud service provider, which is out of the scope of public supervision thus is completely supervised by the cloud service provider. Therefore, the user's personal privacy is not protected, allowing numerous cases where the data being stolen to happen. A large number of users in North America found major websites such as Twitter, Netflix, Paypal, and GitHub inaccessible on October 21, 2016, which lasted six hours [1]. After analysis, we found that, it was due to a DDoS attack from tens of millions of IP which DYN, a well-known DNS service provider that provides domain name resolution services to US Internet companies, suffered. A major source of the attack is the Mirai botnet that has controlled hundreds of thousands of Internet of things devices, including cameras, routers, DVR( hard disk recorders). The number of those devices controlled is still growing at a high speed. March 10, 2021, A security system startup Verkada was hacked and a lot of surveillance footage was stolen. The footage clips of the Tesla Shanghai factory, software provider Cloudflare and other companies monitoring were exposed [2]. Moreover, 150,000 real-time surveillance videos from other hospitals, companies, police stations, prisons, and schools were also exposed in the incident. It is reported that the method used by the hackers was not complicated, they gained access to Verkada through the username and password of an administrator account that was publicly available on the internet, and eventually got inside the Verkada network. On March 15, a group of companies that provide face recognition cameras, which are installed in different kinds of stores, were also named in the CCTV's 315 Gala. They can capture customers' face information without their knowledge and perception, as well as manually add labels to mark various types of customers. IoT surveillance systems with security risks such as this one, can add to the privacy leakage worries of users. Thus, it is the requirement of the era to adopt feasible technical means to protect the privacy of users.

## 1.1 Relevant Work

In addition to surveillance systems, the application areas of IoT are involved in many aspects including intelligent transportation, smart home, public security, etc. For different IoT systems, many scholars have proposed a variety of privacy protection

methods. Jingxue Liao et al. [3] designed a privacy protection system suitable for community Internet of things innovation service platforms by introducing access control and generalization methods. Wang Le et al. [4] improved the traditional CP-ABE scheme to improve the privacy protection efficiency of the wearable health monitoring system. Andre Lizardo et al. [5] introduced a security protocol called Sharelock, which provides end-to-end security and confidentiality for information exchange between communication node groups. The protocol is designed to support the exchange and storage of common messages between nodes that communicate through untrusted edge servers. Almudena Alcaide et al. [6] proposed a fully decentralized anonymous authentication protocol designed to encourage the implementation of privacy-preserving IoT target-driven applications. The system is established by a hoc community of decentralized founding nodes. From then on, nodes can interact and become participants cyber-physical system and maintain complete anonymity.

Home Internet, as a subsystem of smart grid, is usually used in home premises. In the home Internet, household appliances are connected with intelligent meters through wireless connections, and intelligent meters can obtain information such as electricity consumption of electrical appliances from wireless connections. However, due to the insecurity of wireless communication, attackers may intercept and monitor network traffic to obtain the above sensitive information through wireless communication, or even impersonate electrical devices in the home Internet to send forged information to smart meters for benefits. To address the problem of data sources in the home internet, Zhaohui Tang et al. [7] proposed a solution to ensure that the reported energy usage is collected from real devices at designated locations and reflects the actual consumption of devices.

## 1.2 Work on This Paper

In order to avoid the theft of private information recorded by the user's camera and to enhance the information security of the Internet of things systems, a new architecture and scheme of the Internet of things system are proposed. On the basis of the blind decryption algorithm of Kouichi Sakura [8] et al. and the threshold RSA signature algorithm of the Victor Shoup [9], we double encrypt the video information recorded by the camera, and store the private key used for decryption in a secret sharing manner to ensure that no one but the user holding the camera holds complete private key information. Finally, this paper analyzes the security of the proposed scheme from different perspectives.

## 2 Architecture of Internet of Things Surveillance System

Based on the idea of public key encryption, it is assumed that each user holds a pair of public–private key combinations that the user stores the public key in the camera device, while the private key is kept by the user himself. After collecting video information, the camera device encrypts the video information with a randomly generated symmetric key, then uses the public key to encrypt the symmetric key. After these two encryptions, the camera sends the video message ciphertext and symmetric key ciphertext to the cloud service provider. After obtaining the video information ciphertext and the symmetric key ciphertext it held to the cloud service provider. After getting the video information cipher and symmetric key cipher from the cloud service provider, the user only needs to decrypt the symmetric key cipher with the private key and then decrypt the video information cipher with the symmetric key to view the video information. The system architecture at this point is shown in the Fig. 1.

However, when the user can not view the video information at all times while hoping that the video information collected by the camera device can be monitored in real time, a trusted third-party monitoring platform is needed to perform the monitoring task. At the same time, users also demand that the third-party monitoring platform be unable to know their private key. However, the third-party monitoring platform needs the private key to decrypt the symmetric key ciphertext, so a key escrow center is added to our system architecture.

The key escrow center consists of multiple servers which are used to decrypt symmetric key ciphertext with private keys. According to the threshold signature technology, the user shares the private key in secret, thus getting the same number of shared shares as the number of servers in the key escrow center. Next, the user distributes these shared shares to each server in the key escrow center. When the key escrow center receives a symmetric key ciphertext from a third-party monitoring platform, each server it contains will individually sign the symmetric key ciphertext, i.e., decrypt it, and submit the obtained part of the symmetric key to the third-party monitoring platform. After the third-party monitoring platform receives a certain number of partial symmetric keys (threshold  $k$ ), these partial symmetric keys can be combined to obtain the symmetric key. The third party monitoring platform only needs to decrypt the video information ciphertext with the symmetric key to get the video information.

However, if an attacker attacks the key escrow center and obtains a number of partial symmetric keys more than threshold  $k$ , the attacker can also combine the symmetric keys. At this point, the attacker only needs to obtain the video message ciphertext from the cloud server provider and the video information will be leaked. Therefore, according to blind decryption technique, the third-party monitoring platform is required to process the symmetric key ciphertext with a random number before sending the symmetric key ciphertext to the key escrow center. Similarly, after combining the part of the symmetric key processed with a random number, the third-party monitoring platform will then uses the original random number to reverse

the process. In this way, even if the key escrow center decrypts and processes the ciphertext, the symmetric key processed with a random number from the key escrow center that the attacker gets will not be adequate to acquire the symmetric key. The final system architecture is shown in the Fig. 2.

### 3 Encryption Scheme

#### 3.1 Scheme Definition

Based on Sect. 2, a system scheme consists of the following nine algorithms:

**SETUP.** Set RSA module  $n$ , finite domain  $GF(p)$ , open system parameter  $params$ .

**SKE.Encrypt.** Input video information  $M$  collected by camera, randomly generated symmetric key  $mk$ , output video information ciphertext  $C$  encrypted with symmetric key.

**KEM.Encrypt.** Input randomly generated symmetric key  $mk$  and held public key certificate  $e$ , output symmetric key ciphertext encrypted with public key  $ck$ .

**SKE.Decrypt.** Input video information ciphertext  $C$  and symmetric key  $mk$  obtained from cloud server provider, output video information  $M$ .

**KEM.Decrypt.** Input the private key  $d$  and symmetric key ciphertext  $ck$  held by the user and output the symmetric key  $mk$ .

**SecretShare.** Input the private key  $d$  held by the user, the number of servers  $l$  and the threshold  $k$  of the key escrow center, output  $l$  private key secret shares  $ds_j(j = 1, 2, \dots, l)$ .

**RandomProcess.** Input symmetric key ciphertext  $ck$ , random number  $r$  generated by third-party monitoring platform, output processed symmetric key ciphertext  $rck$ .

**ThresholdSign.** Input the processed symmetric key ciphertext  $rck$ , secret shares  $ds_j(j = 1, 2, \dots, l)$  of any  $k$  private keys, output  $k$  partial symmetric key  $rmk_i(i = 1, 2, \dots, k)$  processed by  $r$ .

**CombineSign.** Input public key certificate  $e, \Delta = l!$ , processed symmetric key ciphertext  $rck$ , any  $k$  of  $l$  processed partial symmetric keys  $rmk_i(i = 1, 2, \dots, k)$ , output processed symmetric key plaintext  $rmk$ .

**RandReverseProcess.** Input the processed symmetric key plaintext  $rmk$ , the random number  $r$  generated by the third-party monitoring platform, and output the symmetric key  $mk$ .

#### 3.2 Specific Construction

In this section, video information is encrypted by blind decryption scheme, secret This section constructs a specific video privacy protection scheme using blind decryption

scheme, secret sharing scheme and threshold signature scheme to encrypt video information, which consists of the following 8 algorithms:

**SETUP.** Arbitrarily select two large prime numbers of equal length  $r'$  and  $s'$ , calculate  $r = 2r' + 1, s = 2s' + 1$ , RSA algorithm mod  $n = rs$ , the public key  $e$  is randomly selected, it should be noted that  $e$  must be a prime number so that it satisfies  $\gcd(e, (r-1)(s-1)) = 1$ , calculate the private key  $d = e^{-1} \bmod n$ , construct the finite domain  $GF(p)$ ,  $p = r's'$ , then the system public parameters are  $params = \{n, e, p, GF(p)\}$ .

**KEM.Encrypt.** Given the symmetric key  $mk$  randomly generated by the camera and the public key  $e$  held by the camera, the symmetric key ciphertext encrypted with the public key is  $ck = mk^e \bmod n$ .

**KEM.Decrypt.** Given the private key  $d$  and symmetric key ciphertext  $ck$  held by the user, calculate the symmetric key  $mk = ck^d \bmod n$ .

**SecretShare.** The private key  $d$  held by the user, the number of servers  $l$  in the key escrow center and the threshold  $k$  are given:

- (1) Select  $l$  different nonzero elements  $x_1, x_2, \dots, x_l$  from the finite domain  $GF(p)$ ;
- (2) Choose any  $(k-1)$  elements  $a_i (i = 1, 2, \dots, k-1)$  from  $GF(p)$  to constitute a random polynomial  $f(x)$ :

$$f(x) = d + \sum_{i=1}^{k-1} a_i * x^i \bmod p \quad (1)$$

- (3) Calculate the shares of private key secret sharing:

$$ds_j = f(x_j) = d + \sum_{i=1}^{k-1} a_i * x_j^i \bmod p \quad (j = 1, 2, \dots, l) \quad (2)$$

**RandProcess.** Given the symmetric key ciphertext  $ck$ , the random number  $r$  generated by the third-party monitoring platform and the public key  $e$  held by the camera, the processed symmetric key ciphertext is  $rck = r^e ck \bmod n$ .

**ThresholdSign.** Given the handled symmetric key ciphertext  $rck$ , the shares of private key,  $ds_j (j = 1, 2, \dots, l)$ :

- (1) Define  $\Delta = l!$
- (2) Calculate partially symmetric keys processed by  $r$

$$rmk_j = rck^{2\Delta ds_j} \quad (j = 1, 2, \dots, l) \quad (3)$$

**CombineSign.** Given the public key  $e$ ,  $\Delta = l!$ , the processed symmetric key ciphertext  $rck$ , any  $k$  of  $l$  processed partial symmetric keys,  $rmk_i (i = 1, 2, \dots, k)$ :

- (1) Define  $S = \{i_1, i_2, \dots, i_k\}, \lambda_{i,j}^S = \Delta \cdot \prod_{j' \in S \setminus j} \frac{i-j'}{j-j'}$ , in which  $i \in \{1, 2, \dots, l\} / S, j \in S$

## (2) Combined partial signature

$$\omega = rmk_{i_1}^{2\lambda_{0,i_1}^S} rmk_{i_2}^{2\lambda_{0,i_2}^S} \dots rmk_{i_k}^{2\lambda_{0,i_k}^S} \quad (4)$$

- (3) Get  $a, b$  from the formula  $4\Delta^2a + eb = 1$ , calculate the processed symmetric key plaintext  $rmk = \omega^a rck^b$ .

**RandReverseProcess.** Given the processed symmetric key plaintext  $rmk$ , the random number  $r$  generated by the third-party monitoring platform, the symmetric key is  $mk = rmk/r \bmod n$ .

## 4 Security Analysis

The system architecture scheme proposed by us basically achieves the double encryption of video information, and also satisfies that the two key information of ciphertext and corresponding key are held by different roles in the whole monitoring system. We have implemented the scheme through programming simulation, which also proves that the scheme is correct. In the following, we will analyze the security of the scheme from the perspective of different roles.

**Theorem 1.** If the cloud service provider is controlled by an attacker, the attacker can not obtain or corrupt the video information stored by the user.

**Proof.** It is proved that if the controlled cloud service provider wants to obtain video information, the cloud service provider can only do so in two ways: one is to obtain the private key from the user, which is impossible to achieve because the user will not provide the private key to any other role; the second is to send the symmetric key ciphertext to the key escrow center, which will decrypt the symmetric key ciphertext, which is also impossible to achieve because the key escrow center only provides decryption service to the third-party organization authorized by the user, and the user obviously cannot provide authorization to the cloud service provider, so the key escrow center cannot provide decryption service to the cloud service provider.

**Theorem 2.** When the user cancels the authorization to the third-party monitoring platform, the third-party monitoring platform will not be able to obtain or corrupt the user's video information.

**Proof.** When the user cancels the authorization to the third-party monitoring platform, even if the third-party monitoring platform can obtain video information ciphertext and symmetric key ciphertext from the cloud service provider, the key escrow center still can not provide decryption service to it. Because the key escrow center only provides decryption services to third parties authorized by the user. At the same time, the third-party monitoring platform can not obtain the private key of the user, because the key escrow center does not support the service of providing private key share.

**Theorem 3.** When less than  $k$  servers in key escrow center is compromised by an attacker, the attacker will not be able to access or corrupt the private key information hosted by the user at the key escrow center.

**Proof.** Assuming that the attacker can read or change the share of any private key stored in the server at will, due to the restriction of the secret sharing threshold scheme, if the number of servers maliciously attacked by the attacker in the key escrow center is less than  $k$ , the attacker will not be able to obtain any information about the private key.

**Theorem 4.** If the key escrow center is controlled by an attacker, the attacker can not obtain or destroy the user's video information.

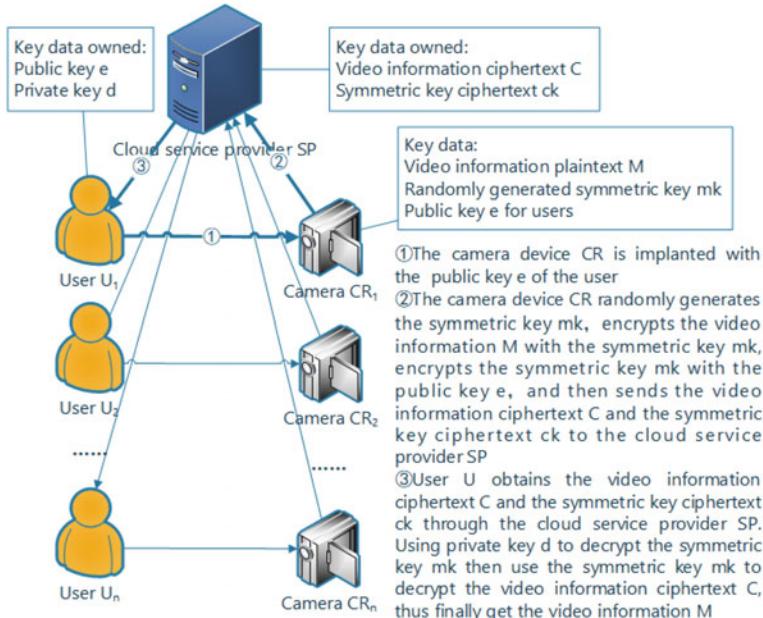
**Proof.** First of all, even if more than or equal to  $k$  servers in key escrow center is compromised by an attacker, the attacker cannot recover the private key according to the secret recovery algorithm, but can only restore the equivalent key

$$d' = f(0) = d \bmod p \quad (5)$$

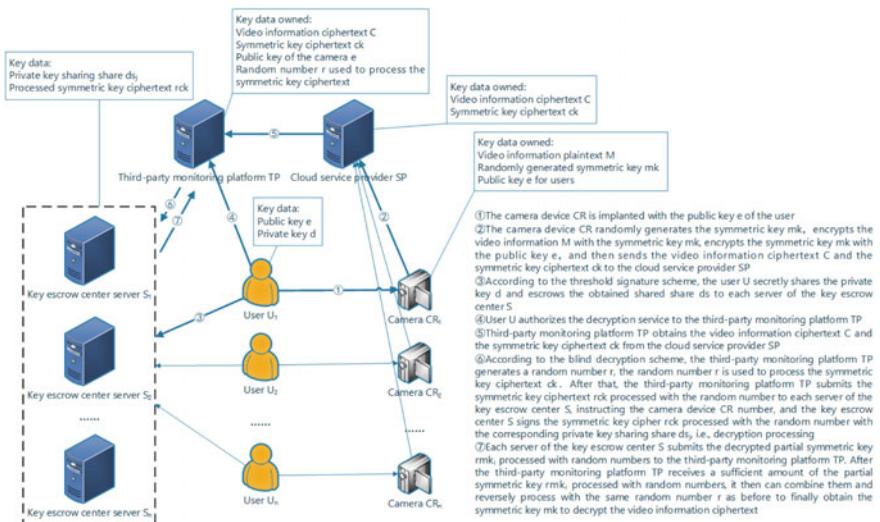
of the private key  $d$ . Since  $p$  is unknown to the key escrow center, the private key is safe in the key escrow center, and the scheme is robust. Secondly, the key escrow center controlled by the attacker cannot obtain the video information ciphertext from it, no matter it is the user or the third-party monitoring platform. Even if the attacker acquires the video message ciphertext from the cloud service provider, the attacker can not obtain the symmetric key plaintext because the symmetric key ciphertext provided by the third party monitoring platform to the key escrow center is processed with a random number held by only it. Even if the key escrow center decrypts, it still obtains the symmetric key plaintext processed with random numbers, so the attacker can not decrypt the video message ciphertext, thus can not obtain the video information.

## 5 Summary

The encryption scheme of the Internet of things monitoring system proposed in this paper achieves the privacy protection of the user's video information through the double encryption of video information and the setting of the key escrow center, which not only effectively prevents theft of users' private information by cloud service providers with ulterior motives but also fills the security loopholes created by irresponsible enterprises due to their negligence of network security. In addition, the encryption scheme proposed in this paper is not only applicable for the monitoring system, but also can be encrypted by other sensors in the Internet of things, which can also realize the privacy protection of users.



**Fig. 1** The proposed IoT monitoring system architecture without key escrow centers and a third-party monitoring platform



**Fig. 2** The proposed IoT monitoring system architecture with key escrow centers and a third-party monitoring platform

## References

1. Internet of Things security gets in trouble frequently—Is the security budget less than 1%? (Chinese), [http://tech.cnr.cn/techgd/20161025/t20161025\\_523219370.shtml](http://tech.cnr.cn/techgd/20161025/t20161025_523219370.shtml). Last accessed 25 Oct 2016
2. 150000 cameras were hacked and Tesla's factory in Shanghai was accidentally exposed! (Chinese), <http://finance.sina.com.cn/tech/csj/2021-03-11/doc-ikkntiak7968500.shtml>. Last accessed 11 Mar 2021
3. Jingxue, L., Fuzhen, C., Jiuju, C., Xiangrong, C.: A privacy protection system for the community IoT innovative technology and service platform. *Netinfo Security* **16**(12), 60–67 (2016)
4. Le, W., Zherong, Y., Rongjing, L., Xiang, W.: A CP-ABE privacy preserving method for wearable devices. *Netinfo Security* **18**(6), 77–84 (2018)
5. Lizardo, A., Barbosa, R., Neves, S., Correia, J., Araujo, F.: End-to-end secure group communication for the Internet of Things. *J. Inf. Secur. Appl.* **58**, 102772 (2021)
6. Alcaide, A., Palomar, E., Montero-Castillo, J., Ribagorda, A.: Anonymous authentication for privacy-preserving IoT target-driven applications. *COMPUT SECUR* **37**, 111–123 (2013)
7. Tang, Z., Keoh, S. L.: An Efficient Scheme to Secure Data Provenance in Home Area Networks. pp. 115–120. IEEE (2020)
8. Sakurai, K., Yamane, Y.: Blind decoding, blind undeniable signatures, and their applications to privacy protection. pp. 257–264. (Springer, Berlin, Heidelberg, 1996)
9. Shoup, V.: Practical Threshold Signatures. pp. 207–220 (Springer, Berlin, Heidelberg, 2000)

# An Overview of Text Steganalysis



Yu Yang , Lei Zha , Ziwei Zhang , and Juan Wen

**Abstract** With the rapid development of the Internet, network information security is progressively under menace. Text steganography is one of the key reasons to affect information content security. It aims to hide confidential information in text carriers in a concealment system. In case that text steganography is used by criminals, malicious information can easily be transmitted through the Internet without being discovered by a third party. In contrast, text steganalysis is an effective technique to solve this problem by detecting whether a text carrier contains secret information. This paper presents an overview of current text steganalysis methods starting from 2006. We discuss the basic text steganalysis model and compare the pros and cons of these algorithms, hoping to offer some perceptions and motivations for future research directions.

**Keywords** Information security · Text steganography · Text steganalysis

## 1 Introduction

With the speedy development of information technology, the unseal surroundings of information transmission and sharing has been rapidly constructed. While providing convenience for people's lifestyle, it brings a series of safety risks. For example, the network information is susceptible to spiteful assaults, illegal access, falsification, plagiarism, etc. [1]. How to ensure the safety of multimedia data has become a significant topic that needs to be resolved eagerly in the domain of information security.

Steganography is the aesthetics and technology to conceal confidential information into multimedia carriers. Modern steganography technology uses mankind perception redundancy, statistical redundancy of multimedia data, and other characteristics to hide secret information by a certain coding form or encryption in some

---

Y. Yang · L. Zha · Z. Zhang · J. Wen ()

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

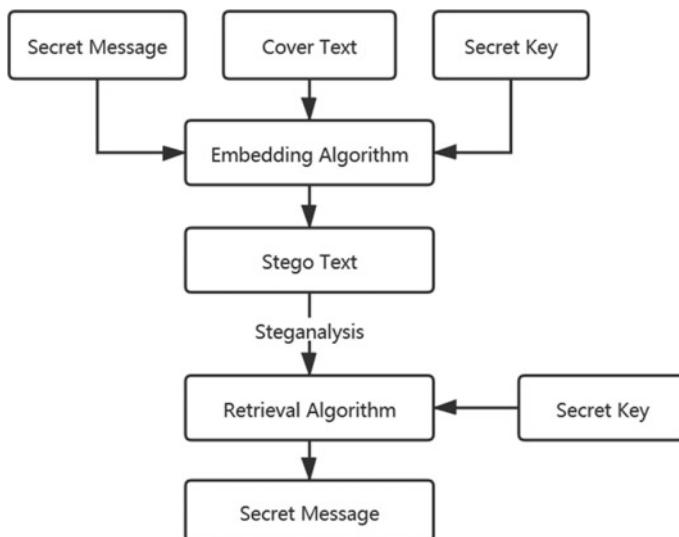
e-mail: [wenjuan@cau.edu.cn](mailto:wenjuan@cau.edu.cn)

public data carrier to embed confidential information. The cover used to hide classified message is generally multi-media files transmitted on the network, such as videos, audio, images, text, etc. Among them, text has become an important cover due to its fast transmission and convenient access. Particularly, with the rapid development of natural language processing (NLP), text steganography has been greatly developed and refined. Currently, text steganography has been extensively used in confidential correspondence, copyright maintenance, content identification, etc.

Unlike text steganography, text steganalysis identifies whether a provided text contains undisclosed communication and extracts the embedded secret information when possible. In recent years, text steganalysis is becoming a vital investigation topic on information security, as one of the effective ways to prevent criminals from malicious use of text steganography technology for illegal activities. In addition, it further ensures safe and covert communication, and has important applications in military, intelligence, and government secret departments, such as detecting and jamming enemy communication signals; it can effectively block information sources, and conduct information reconnaissance and destruction on the enemy. Almost all information embedding algorithms inevitably change the statistical characteristics of the carrier. The core idea of steganalysis is to use a statistical machine learning algorithm to model and detect the subtle differences caused by information embedding, so as to identify the suspicious and stego text.

The key of text steganography and text steganalysis is shown in Fig. 1.

Owing to the significance of text steganalysis in Internet security, it is essential to review and summarize the mainstream text steganalysis in recent years. This



**Fig. 1** Basic model of text steganography and text steganalysis [2]

paper outlines the current state of research in text steganalysis starting from 2016. Furthermore, we summarize, compare and analyse some of these algorithms.

Next, introduce the framework of this article. In Sect. 2, different types of text steganalysis are reviewed. The techniques and concepts involved in each type of text steganalysis are described in detail. In Sect. 3, a comparative analysis of the techniques and approaches is made. Eventually, the conclusion is drawn in Sect. 4.

## 2 Classification of Text Steganalysis Algorithms

In this Section, the classification of text steganalysis, including targeted steganalysis and blind steganalysis, will be introduced in detail.

### 2.1 Targeted Text Steganalysis

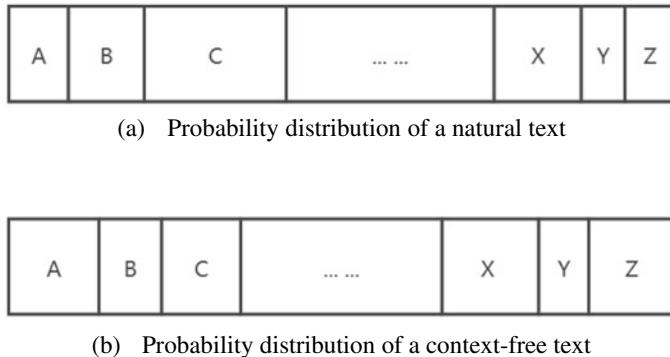
Targeted text steganalysis is a steganalysis means introduced to identify an especially text steganography algorithm. Scilicet, the detection algorithm knows which text steganography method is used to embed the secret information. Thus, targeted steganalysis is excel in detecting the specific text steganography algorithm. However, they may fail exponentially when they facing other steganography algorithms.

The common statistical features used for targeted steganalysis include word-initial distribution, alphabetic cases, contextual information, evolutionary features, and synonym frequency.

**Distribution of First Letters of Words** [3]. For the stego text generated by context-free steganography, words occur randomly, and the probability of appearance of words in each segment of the text lies only on the possibility in local region. In contrast, in a natural text, words do not occur randomly, and the process of word generation can be viewed as an nth-order Markov process [2]. That is, the probability distribution of word initials in natural texts is very different from that of word initials in context-free texts, as shown in Fig. 2.

**Stego** [4]. Stego is a text steganography tool that uses dictionaries to transform secret message into grammar-free text with a configuration similar to normal text for steganographic communication. By studying the mechanism of Stego, the paper [4] proposes a Stego-based text steganography analysis method. When the dictionary words used for steganography start with all lowercase letters, the stego text can be detected by the steganalysis method based on sign features. Otherwise, the stego text will be detected by the steganalysis method based on statistical features.

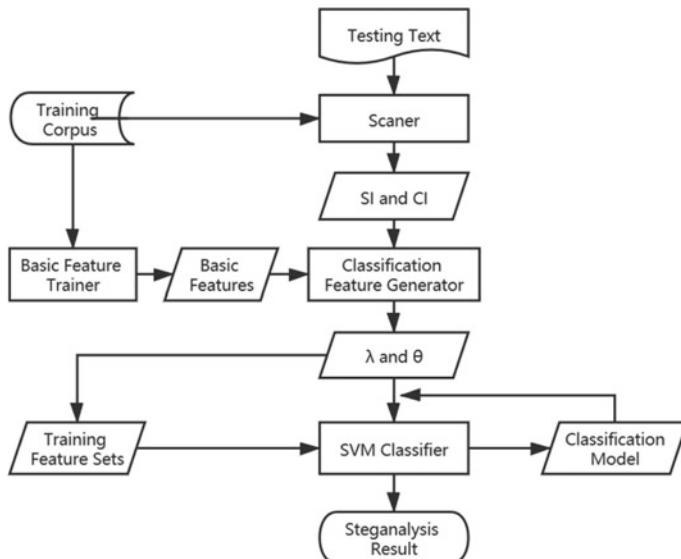
**Context Information.** Article [5] introduces the concept of context clustering to estimate the contextual fitness of a text and shows how to distinguish ordinary text from a stego text by counting the contextual fitness values of the text. The Substitution-based Linguistic Steganography (SLS) system replaces an original element in the overwritten text with a replacement element in the same replacement



**Fig. 2** Distribution of probabilities of natural text and context-free text

set when performing message steganography. This substitution behaviour may result in the new replacement element not fitting well to the original context. According to this feature, the paper proposes a steganalysis scheme for SLS, and the specific process is shown in Fig. 3. Following that, a text steganalysis method based on synonym replacement is proposed based on this scheme. The average accuracy of this text steganalysis approach is 98.86%.

Article [6] proposes a word embedding-based approach to detect secret information in a text. The method uses a continuous Skip-gram model to symbolize synonyms



**Fig. 3** The steganalysis direct at substitution-based text steganography [5]. SI: Substitution Information; CI: Context Information;  $\lambda$ : Context Maximum Rate;  $\theta$ : Context Maximum Deviation

and their contextual words as word embeddings and encode the word semantics as a low-dimensional dense vector; the embeddings of synonym counterparts are used to effectively estimate the contextual adaptation and are weighted by the TF-IDF scores of the contextual words. By analysing the distinctions in the contextual adaptation scores of synonyms in the synonym set and the distinctions in the contextual adapt values of synonyms in the cover text and the stego text, extract three features and then input them to a support vector machine (SVM) classifier for steganalysis. The proposed steganalysis technique enhances higher than 4.8%.

**Evolution Algorithm** [7]. Article [7] proposes an evolutionary detection steganalysis system (EDSS) based on the evolutionary algorithm of the Java Genetic Algorithm Package (JGAP). The results of the EDSS can be classified into good adaptation and bad adaptation according to the adaptation value.

**Synonym Frequency** [8]. Article [8] proposes a text steganalysis method based on synonym substitution (SS). First, attribute pairs of synonyms are introduced to represent their positions in the ordered synonym set and the size of synonyms. Due to the substitution of synonyms, the quantity of high-frequency attribute pairs decreases nevertheless the quantity of low-frequency attribute pairs increases. Ground on this, the changes of statistical features of SS steganographic pairs of attributes are analysed theoretically, and secret information is detected using eigenvectors build on the relative frequency differences of diverse attribute pairs. This paper also analyses the impact of the synonym encoding strategy on feature vector extraction.

## 2.2 *Blind Text Steganalysis*

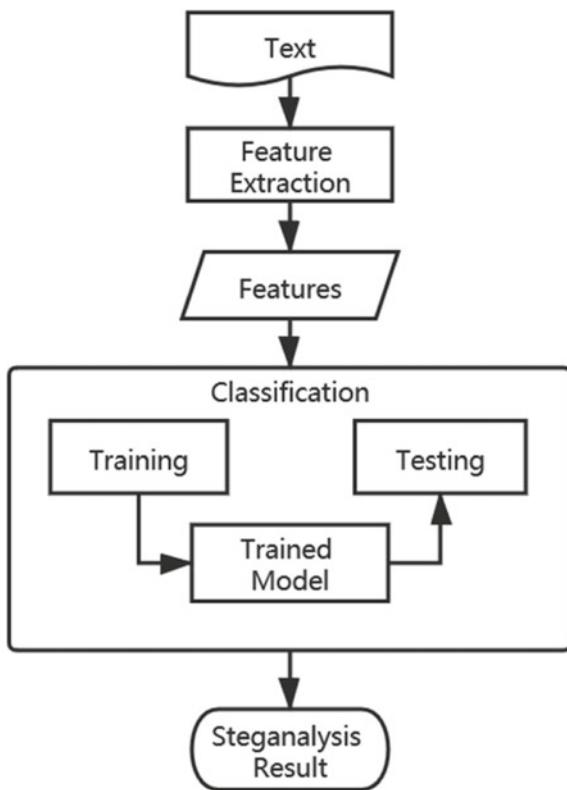
Blind steganalysis does not depend on a specific steganographic algorithm. As a result, it meets a wider range of applications and requirements. Since embedding secret information in normal text more or less changes the content of the text, introducing statistical difference in normal textual features. Therefore, the key step for blind steganalysis is to model these subtle differences [9]. As Fig. 4 shows, feature extraction and text classification are two stages of blind text steganalysis. Next, we will introduce the current mainstream blind steganography algorithms based on different model types.

**Text Steganalysis Based on AdaBoost** [10]. It points out that the statistical changes will be brought to the text after embedding secret information. Ground on this, a general detecting algorithm ground on AdaBoost is put forward to extract text statistical features and detect natural texts and stego texts.

AdaBoost can recognize all text embedding rates at 2 and 4%, and the recognition rate is also 100% under other conditions. The experiment proves that AdaBoost is almost unaffected by the embedding rate, reflecting the superior classification performance of AdaBoost.

**Text Steganalysis Based on Statistical Language Model** [11]. In the article [11], a text steganalysis algorithm based on a statistical language model is proposed to classify a given text segment into natural text and stego text using its complexity. The

**Fig. 4** Standard blind text steganalysis phases



algorithm achieves 96.3% recognition accuracy for stego text segments and natural text segments when the segment size is 5 K; the algorithm detects more than 93.9% accuracy when the text size is 2 K. Not only that, but the experiment also tested the NICETEXT system, TEXTO system, and the text generated based on the Markov chain, and achieved superior results.

**Text Steganalysis Based on SVM [12].** Article [12] proposes an SVM-based hidden information detection algorithm. The SVM classifier is built by learning and training the normal text and small-sample laden confidential text, and the better generalization ability of the classifier is used to classify the unknown text. The model has great generalization performance and the SVM classifier also has an excellent classification effect.

**Natural Frequency Zoned Word Distribution Analysis (NFZ-WDA) [13].** Translation-based steganography (TBS) is secure text steganography that encodes secret information using the noise generated by the translation of natural language text. The NFZ-WDA method proposed in article [13] aims to detect TBS without using any TBS-related information. The single support in this method is a natural frequency lexicon, a word frequency dictionary obtained from a large corpus. NFZ-WDA uses frequency criteria (NFZs) to refine word distribution features. Since the

elaboration of word distribution features maintains more structural information, the improved method can analyse the stego text generated by TBS more effectively. To attest the validity of the NFZ-WDA method, the paper carries out experiments on two-class and multi-class SVM classifiers. The results show that the accuracy of both detections is comparatively high and increases with the increase of text size. Thus, this text steganalysis method has good application prospects.

**Text Steganalysis Based on Convolutional Neural Network (CNN).** Article [14] proposes a CNN-based model for text steganalysis that captures complex dependencies and automatically learns the text feature representations. A decision strategy for detecting long texts is also proposed, so as to boost the performance ulteriorly. Firstly, the word embedding layer extracts the semantic and syntactic features of words. Secondly, use different sized rectangular convolution kernels to learn sentence features. The method is not only valid in exploring different types of text steganography algorithms but also achieves excellent results in analysing texts of different sizes.

Article [15] propounds a two-stage CNN-based method for text steganalysis. The first stage is a sentence-level CNN, consisting of a convolutional layer containing multiple convolutional kernels with disparate window sizes, a pooling layer, a fully connected layer with Dropout, and a Soft-max output. In this way, the layer not only handles variable-length sentences but also obtains two steganographic features per sentence. The second stage is a text-level CNN that uses the output of the first stage to ensure whether the detected text is steganographic or not. The average accuracy of this approach is 82.245%.

**Text Steganalysis Based on Recurrent Neural Networks (RNN)** [16]. In automatically generated stego text, the distortion of the conditional probability distribution is caused by the embedding of hidden information. Based on this, paper [16] proposes a text steganalysis algorithm that uses RNN to extract these feature distribution differences and subsequently classify these features into cover text and steganographic text. The experimental results show that the model not only has high detection accuracy but also can use the subtle differences of text feature distributions to estimate the amount of information embedded in the generated stego text.

**Text Steganalysis Based on Word2vec** [17]. A Word2vec-based approach to text steganalysis is proposed in [17]. First, a multi-dimensional word vector containing rich semantic information is trained for each word using the distributed word representation tool Word2vec; then to calculate the suitability of the synonym in a particular context, the correlation between two words needs to be measured by the cosine distance between the synonym and its contextual word vector, and obtain detection features; finally, the extracted detection features are input into a Bayesian estimation model for training and testing. The average detection accuracy of the approach reaches 97.71% for stego texts with different embedding rates, which has a very good measuring performance.

**Text Steganalysis Based on Convolutional Sliding Windows (TS-CSW)** [18]. Word association features in the stego text are distorted after inserting confidential message, and the TS-CSW is proposed based on this changed feature, which uses convolutional sliding windows (CSW) of multiple sizes to obtain relevant features of

the text. Samples collected from the T-Steg dataset are used in the paper to train and test the proposed steganalysis approach. The model not only has great performance in steganalysis but also can estimate the amount of secret information embedded in the stego text.

**Text Steganalysis Based on Long Short-Term Memory Networks (LSTM) [19].** To enhance the low-level features in the feature vector and then better associate with the low-level features to test the steganographic information in the generated text, paper [19] introduces two parts, including dense connectivity and feature pyramid. It comes up with a text steganalysis approach ground on densely connected long short-term memory networks with a feature pyramid. Firstly, map the words in the text to a semantical space with hidden representations for better utilization of semantical features; then the semantic features at different levels are extracted using a stacked bidirectional long short-term memory networks (Bi-LSTM); finally, fuse the semantic features at all levels and use the Sigmoid layer to resolve whether the text is steganographic or not. This approach achieves a satisfying result.

**Text Steganalysis Based on LSTM-CNN.** In article [20], a hybrid text steganalysis method (R-BILSTM-C) is proposed by combining the advantages of Bi-LSTM and CNN. The method captures long-term semantic information of text using Bi-LSTM and extracts local relationships between words using asymmetric convolutional kernels of different sizes. The detection accuracy is extremely increased. Furthermore, the paper visualizes the high-dimensional semantic feature space. The approach is able to be effectually used to different text steganography algorithms.

Article [21] proposes an LSTM-CNN model for text steganalysis. Firstly, map the words to semantical space to better utilize the semantical features of the text; then LSTM and CNN are combined to obtain local contextual info and long-range contextual info in a stego text. In addition, the text also employs an attention mechanism to identify important cues in suspicious sentences. The model can accomplish outstanding results in steganalysis tasks.

**Text Steganalysis Based on Bi-LSTM-GNN [22].** A text steganalysis model with two stages of high robustness is proposed. In the first phase, Bi-LSTM is used to obtain feature information of all words in a sentence while holding a powerful correlation. In the second phase, input multi-sentence vectors to graph neural network (GNN), from which anomalous features between sentences are extracted. Moreover, article [22] adds adversarial instances to the training set to increase the robustness and generalization of the steganalysis model. The experiments reveal that the model not has excellent robustness but is quite effective for steganographic text judgment.

**Text Steganalysis Based on Capsule Network [23].** Capsule networks identify the subtle differences between stego texts and normal texts by extracting and preserving the semantic features of the texts. Article [23] uses capsule networks to detect whether the natural text contains secret information: the text is vectorized using word2vec, and steganographic text generated by RNNs and variable-length encoding is used as the experimental dataset to enhance the generalization of the method. Experiments reveal that the method can reach a 92% correct detection rate for stego text at a lower embedding rate (1–3 bits/word), which is about 7% better

than that of other neural networks; at a high embedding rate (4–5 bits/word), the detection accuracy can reach more than 94%.

### 3 Evaluation

From the above overview of text steganalysis in the past decade, it can be seen that the development of text steganalysis is consistently changing and improving, from the early target steganalysis to the more versatile and effective blind steganalysis.

The advantages and disadvantages of five chosen target text steganalysis are listed in Table 1. From Table 1, it is clear that the algorithms based on initial letter probability distribution, contextual information, and synonym frequency algorithms are simple and efficient; among them, the contextual information approach is simpler and easier to implement than the other two methods. The Stego-based steganalysis algorithm, however, relies on detecting the case form of the initial letter of text words, which is more restrictive.

As for blind text steganalysis, start from the CNN-based text steganalysis algorithm in [14], it has continuously developed and improved. As can be seen from Sect. 2.2 of this paper, blind steganalysis have been getting better from the early use of machine learning algorithms, such as SVM, to the use of deep learning algorithms such as CNN, RNN, LSTM, and the combination of LSTMs, CNN, and GNN, which have emerged in the last two years. The average detection accuracies of blind text steganalysis for stego texts are listed in Table 2. Although deep learning enhances the property of text steganalysis, the computation complexity and time cost of the algorithm are also rising, which has become one of the issues to be solved in the future.

**Table 1** Comparative analysis of target text steganalysis [9]

No	Years	Methods	Advantages	Disadvantages
1	2006	Distribution of first letters of words	High recall and low error	Require much time
2	2006	Stego	Simple	Require distribution of first letters
3	2011	Context information	Simple and effective variants	Lack of vocabulary
4	2014	Evolution algorithm	Support the text-based document	Complexity of computation
5	2018	Synonym frequency	High speed	Complex

**Table 2** Average detection accuracies of blind text steganalysis

No	Years	Methods	Accuracies
1	2007	Text steganalysis based on AdaBoost	100%
2	2009	Text steganalysis based on statistical language model	Higher than 93.90%
3	2009	Text steganalysis based on SVM	89.80%
4	2011	NFZ-WDA	Higher than 91.22%
5	2019	Text steganalysis based on CNN	82.25%
6	2019	Text steganalysis based on RNN	Higher than 90%
7	2019	Text steganalysis based on Word2vec	97.71%
8	2020	TS-CSW	Higher than 90%
9	2020	Text steganalysis based on LSTM	90.61%
10	2020	Text steganalysis based on LSTM-CNN	91.35%
11	2020	Text steganalysis based on Bi-LSTM-GNN	Higher accuracy
12	2021	Text steganalysis based on capsule network	92% (1–3 bits/word) 94% (4–5 bits/word)

## 4 Conclusion

This paper reviews different types of text steganalysis algorithms since 2006, including target steganalysis and blind steganalysis, and compares and analyses the two categories, respectively. The study indicates that steganalysis methods do have their own advantages and disadvantages. We believe this paper can supply motivation and assistance for future steganalysis research.

As far as the current research trends are concerned, the development of NLP has a significant impact on text steganography and text steganalysis, for most of the latest algorithms are inspired by the advanced technology in NLP. The most momentous issue of text steganalysis is to enhance the effectiveness and robustness of steganalysis while simplifying model complexity. Therefore, in the near future, based on clarifying the development of text steganalysis and its actual development, we will face its main problems, closely combine the latest research results of NLP, reinvent the text steganalysis method, and strive to break through the development bottleneck mentioned in the previous section.

## References

1. Ahvanooy, M., Li, Q., Hou, J., Rajput, A.R., Chen, Y.: Modern text hiding, text steganalysis, and applications: a comparative analysis. *Entropy* **21**, 355 (2019)
2. Chang, C., Clark, S.: Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. *Comput. Linguist.* **40**, 404–448 (2014)
3. Sui, X., Luo, H., Zhu, Z.: A steganalysis method based on the distribution of first letters of words. *IEEE Comput. Soc.* **6**, 369–372 (2006)

4. Wu, M., Jin, S.: Text steganalysis method—breaking steganographic utility of Stego. *Computer Eng.* **32**, 10–12 (2006)
5. Chen, Z., Huang, L., Miao, H., Yang, W., Meng, P.: Steganalysis against substitution-based linguistic steganography based on context clusters. *Comput. Electr. Eng.* **37**, 1071–1081 (2011)
6. Xiang, L., Yu, J., Yang, C., Zeng, D., Shen, X.: A word-embedding-based steganalysis method for linguistic steganography via synonym substitution. In: 6th IEEE Access, pp. 64131–64141 (2018)
7. Puriwat, L.: A detection method for text steganalysis using evolution algorithm (EA) approach. *Adv. Comput. Sci.*, pp. 22–23 (2012)
8. Xiang, L., Sun, X., Luo, G., Xia, B.: Linguistic steganalysis using the features derived from synonym frequency. *Multimed. Tools Appl.* **71**, 1893–1911 (2014)
9. Lokman, S., Mustapha, A., Ismail, A., Din, R.: Analysis review on linguistic steganalysis. *Indones. J. Electr. Eng. Comput. Sci.* **17**, 950–956 (2019)
10. Sui, X., Shen, L., Yan, J., Zhu, Z.: Text steganalysis using AdaBoost. *Tongxin Xuebao* **28** (2007)
11. Meng, P., Hang, L., Yang, W., Chen, Z., Zheng, H.: Linguistic steganography detection algorithm using statistical language model. *Technol. Comput. Sci.* **2**, 540–543 (2009)
12. Xin, G., Hui, L., Zhong, Z.: Text steganalysis based on support vector machine. *Comput. Eng.* **35**, 188–191 (2009)
13. Chen, Z., Huang, L., Meng, P., Yang, W.: Blind linguistic steganalysis against translation based steganography. *Lect. Notes Comput. Sci.* **6526**, 251–265 (2011)
14. Wen, J., Zhou, X., Zhong, P., Xue, Y.: Convolutional neural network based text steganalysis. *IEEE Signal Process. Lett.* **26**, 460–464 (2019)
15. Xiang, L., Guo, G., Yu, J., Sheng, V., Yang, P.: A convolutional neural network-based linguistic steganalysis for synonym substitution steganography. *Math. Biosci. Eng.* **17**, 1041–1058 (2020)
16. Yang, Z., Wang, K., Li, J., Huang, Y., Zhang, Y.: TS-RNN: text steganalysis based on recurrent neural networks. *IEEE Signal Process. Lett.* **26**, 1743–1747 (2019)
17. Yu, J., Xiang, L., Zeng, D.: Natural language steganalysis method based on Word2vec. *Comput. Eng.* **45**, 309–314 (2019)
18. Yang, Z., Huang, Y., Zhang, Y.: TS-CSW: text steganalysis and hidden capacity estimation based on convolutional sliding windows. *Multimed. Tools Appl.* **79**, 18293–18316 (2020)
19. Li, H., Jin, S.: Text steganalysis based on capsule network with dynamic routing. *IETE Tech. Rev.* **38**, 72–81 (2021)
20. Yang, H.: Linguistic steganalysis via densely connected LSTM with feature Pyramid. 2020 Assoc. Comput. Mach. **20**, 5–10 (2020)
21. Niu, Y., Wen, J., Zhong, P., Xue, Y.: A hybrid R-BILSTM-C neural network based text steganalysis. *IEEE Signal Process. Lett.* **26**, 1907–1911 (2019)
22. Bao, Y., Yang, H., Yang, Z., Liu, S., Huang, Y.: Text steganalysis with attentional L STM-CNN. In: 5th Int. Conf. Comput. Commun. Syst., pp. 138–142 (2020)
23. Li, E., Fu, Z., Chen, S., Chen, J.: A two-stage highly robust text steganalysis model. *J. Cyber Secur.* **2**, 183–190 (2020)

# Design and Experiment of UAV Variable Spray Control System Based on RBF-PID



Yunling Liu , Yan Ma , Bowen Wu , and Yajia Liu

**Abstract** Due to the problem of uneven spray deposition and distribution caused by the change of flight speed of plant protection UAV and the accuracy and stability of the UAV variable spray system, a variable spray control system based on RBF-PID was designed in this paper. The system uses GPS and accelerometer to obtain the flight speed of UAV. The relationship between duty ratio and spray flow was established through experiments, on this basis, the spray flow was regulated by the combination of RBF and PID. The simulation results show that the dynamic performance of RBF-PID is better than that of PID and fuzzy PID. Compared with PID and fuzzy PID, the settling time of RBF-PID is shortened by 0.44 s and 0.34 s respectively, and the overshoot is 6.1%, which is less than 27.7% of PID and 25.32% of fuzzy PID. The feasibility of the system is verified by control accuracy experiments and response speed test, and the results show that within the speed range of 2–6 m/s, the relative error between the actual spray flow and the desired flow does not exceed 10%, and the response speed is less than 300 ms under different set flow.

**Keywords** Plant protection UAV · Flight speed · Variable spray · PWM · RBF-PID

## 1 Introduction

For a long time, China has been constrained by backward pesticide spray application equipment and technology, and the pesticide utilization rate has been at a low level. In order to solve the problems of low spraying efficiency and heavy pollution caused by traditional spraying methods, accurate pesticide spraying equipment are required [1]. Variable spray technology is an important way to achieve precise pesticide application [2, 3], it can automatically adjust the spray amount according to the speed [1, 4],

---

Y. Liu · Y. Ma · B. Wu

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

Y. Liu ()

College of Science, China Agricultural University, Beijing 100083, China

e-mail: [liuyajia@cau.edu.cn](mailto:liuyajia@cau.edu.cn)

crop density and hazard degree, thereby reducing pesticide residues in crops and improving the utilization rate of pesticides [5, 6]. At the same time, it can effectively solve the problem of uneven pesticide deposition distribution caused by the change of flight speed in practical operation.

A single sensor is usually used to obtain the real-time velocity of UAV (Unmanned Aerial Vehicle) in the process of variable spraying. Liu et al. [4] used GPS to detect the position of UAV, and calculated the real-time flight speed according to the change of coordinates per unit time, but the deviation level of speed measurement at low speed remains to be verified. Cen et al. [7] designed a PID (Proportion Integration Differentiation) spray system based on neural network, and used air pressure transmitter to get the speed of UAV, but did not explain the error level of speed acquisition. Peng [8] proposed a UAV speed prediction technology based on BP (Back Propagation) neural network, which reduced the speed prediction error to less than 0.1 m/s, but the structure of the algorithm was more complex, and had higher requirements for the performance of the controller.

The research on the control method of variable spray system can be divided into ground equipment and aviation equipment according to the scene. Lebeau et al. [9] developed a PWM (Pulse-Width Modulation) based spray controller, which compensated for the effect of horizontal boom movement velocity on spray deposition uniformity. Wen et al. [10] developed a PID based variable width PWM spray system, which realized variable spray operation under different spray requirements. Fritz et al. [11] used wind tunnel to simulate the scene of agricultural aircraft flight, the effect of different flight speeds on the spray effect was measured. Sun et al. [12] designed a PID control variable spraying system based on neural network tuning, which solved the problems of large steady-state error and long response time of the existing variable spray control algorithm.

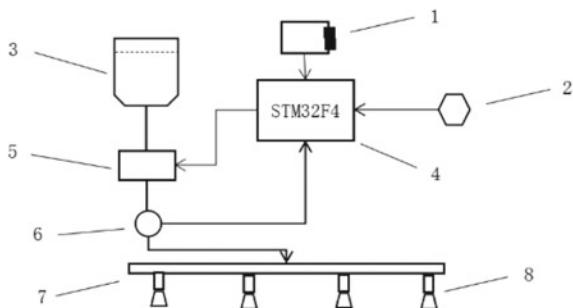
This study focuses on two problems: (1) The real-time flight speed of UAV cannot be accurately obtained by GPS when it is in the acceleration and deceleration stage during spraying operation. (2) During the spraying process, when the velocity of UAV changes, the spraying deposition will be uneven. In this study, speed acquisition is achieved by combining GPS and accelerometer. On this basis, an UAV variable spray control system based on RBF-PID control method is designed, and the effectiveness of the system is proved through simulation experiments, control accuracy experiments and response speed test.

## 2 Design of UAV Variable Spray System

### 2.1 Overall System Design

The UAV variable spray system based on RBF-PID was mainly composed of a GPS sensor, an accelerometer, a flowmeter, an embedded single-chip computer, electronic

**Fig. 1** Schematic diagram of the structure of the variable spray system. Note 1. Power supply; 2. GPS; 3. Liquid tank; 4. Embedded single-chip computer; 5. Electronic governors and micro diaphragm pump; 6. Flowmeter; 7. Spray rod; 8. Nozzle



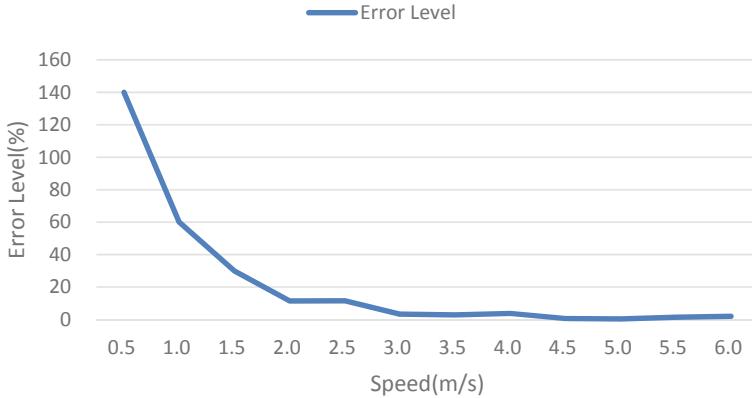
governors and diaphragm pumps, nozzles, etc. The control structure diagram of this system is shown in Fig. 1.

In the system, GPS and accelerometer were selected to obtain the velocity data of UAV in real time, then the desired flow was calculated according to the speed. The real-time pulse number of liquid spraying was obtained by flowmeter, then the pulse number per second was converted into actual flow and transmitted to variable spray controller for processing. The variable spray controller compared the actual flow with the desired flow, continuously regulated the flow deviation through the RBF-PID control algorithm, and outputted the control signal to electronic governors, which controlled the speed of the diaphragm pump, so as to change the actual flow in the system, and finally sprayed the liquid through the high-pressure atomizing fan nozzles. All hardware parts of the whole system were connected by 10 mm hose.

## 2.2 Velocity Acquisition Method Based on GPS and Accelerometer

In order to control the cost of the system, we select common GPS and accelerometer on the market. GPS has the problem of large error between the measured speed and the actual value at low speed, and the accelerometer has the problem of error accumulation. Therefore, this paper develops a method using the accelerometer to correct the GPS measured data, so as to solve the problem that GPS cannot accurately obtain the speed value when UAV is in acceleration or deceleration state.

In order to determine the speed range where the GPS requires data correction with an accelerometer, GPS was fixed on the small vehicle. Then the vehicle was driven at different speeds. Both the speed of the vehicle and the speed measured by GPS were recorded, and the errors of GPS at different speeds were calculated. The flight speed of the plant protection UAV during operation is usually 3–6 m/s, generally not more than 6 m/s, so the speed test range is set as 0.5–6 m/s, and the results are shown in Fig. 2. When the speed is within the range of 3–6 m/s, the error of the speed measured by GPS is not more than 4%, which means the GPS speed can be used as the actual speed at this time. When the speed of the vehicle is less than 3 m/s, the error of the



**Fig. 2** Error level of GPS module at different speeds

speed value measured by GPS is more than 11.50%, it means that GPS speed needs to be corrected. So, this paper takes 3 m/s as the critical value of whether the velocity measured by GPS needs to be corrected.

Based on the above test results, this paper proposes a method of UAV speed acquisition which combines GPS and accelerometer: when the flight speed measured by GPS is higher than 3 m/s, accept it as the real speed. When the speed measured by GPS is less than 3 m/s, the current speed value is taken as the initial value, the corrected real-time speed will be the sum of the initial value and the integral result of the acceleration value measured by the accelerometer. The equation of the real-time speed can be defined as:

$$V = V_0 + \int a dt \quad (1)$$

where  $V$  is the real-time speed of UAV,  $V_0$  is the speed when the UAV enters the deceleration stage, and  $a$  is the acceleration value of UAV forward direction measured by accelerometer.

### 2.3 Control Method of Spray System

**Relationship Model between Speed and Flow.** Firstly, the relationship model between UAV flight speed  $V$  and spray flow  $Q_{rin}$  is introduced as the basis for calculating desired flow, which can be defined as follows [13]:

$$Q_{rin} = \frac{Nvd}{166.67} \quad (2)$$

where  $Q_{rin}$  is desired flow in L/min,  $N$  is the pesticide application volume per unit area in L/hm<sup>2</sup>,  $v$  is the flight speed in m/s, and  $d$  is the spray swath in m.

**Relationship Model between Duty Ratio and Spray Flow.** It is necessary to calibrate the spray flow of the micro diaphragm pump with PWM wave duty ratio [12]. When the duty ratio is less than 45%, the diaphragm pump cannot work, and when the duty ratio is higher than 80%, the spray flow will no longer change. Therefore, the duty ratio range is set to 45~80% and the step length is 1%. The relationship between the spray flow and the duty ratio of the PWM wave is obtained. The result is shown in Fig. 3.

The relationship between duty ratio and spray flow can be obtained by 3 polynomial fitting equation, which can be expressed as:

$$y = -3 \times 10^{-5}x^3 + 0.0053x^2 - 0.2407x + 3.271 \quad (3)$$

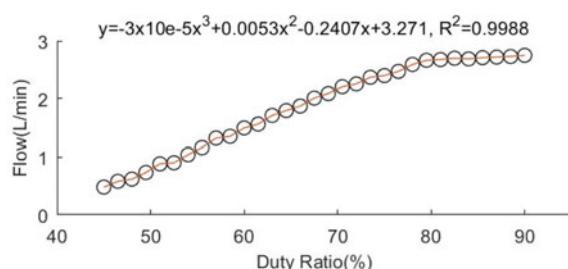
where  $y$  is actual spray flow,  $x$  is the duty ratio, and the coefficient of determination ( $R^2$ ) of the model is 0.9988, which indicates that the model has high fitting degree.

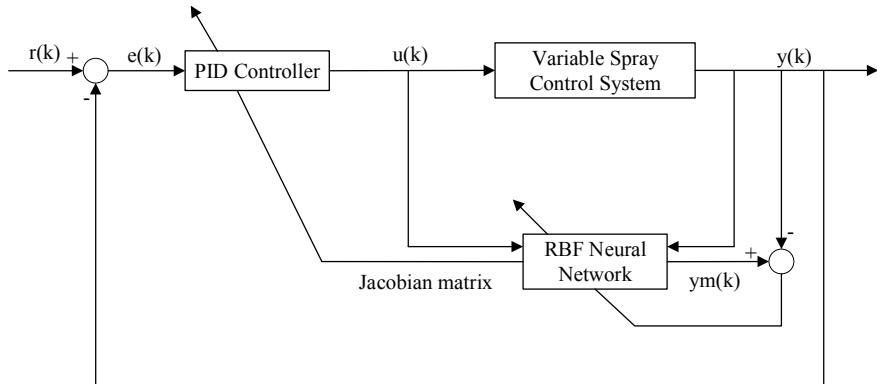
**RBF-PID Control Method.** In this paper, RBF-PID method is used for variable spraying control. This method combines RBF neural network with PID control method, and adjusts the parameters of PID controller through RBF neural network. RBF neural network is used as identifier to identify PID controller, and the parameters of PID controller are adjusted according to the identification results. Li et al. [14] and Zeng et al. [15] used hybrid control strategy combining RBF and PID to control UAV attitude and greenhouse climate in simulation conditions. In this paper, the control principle of variable spray system based on RBF-PID is shown in Fig. 4 [16, 17].

The workflow of the system is as follows:

- (1) The flight speed of the UAV is obtained, and the variable spray control system calculates the desired spray flow according to Eq. (2) as the desired flow  $r(k)$  of the control system.
- (2) The deviation  $e(k)$  between the actual spray flow and the desired flow is calculated, the deviation is used as the input of the PID controller, and the PID controller outputs the control signal  $u(k)$ .

**Fig. 3** Relationship between duty ratio and spray flow





**Fig. 4** Control schematic diagram of RBF-PID

- (3)  $u(k)$  is transmitted to the variable spraying control system, and the duty cycle of PWM is changed to adjust the actual output flow of the system  $y(k)$ .
- (4) The Jacobian information is obtained by RBF neural network identification. The Jacobian information is used as the basis to adjust the parameters of PID controller, so as to realize the variable application of parameter self-tuning.

### 3 Simulation Experiment Design and Result Analysis

#### 3.1 Simulation Experiment of Control Method

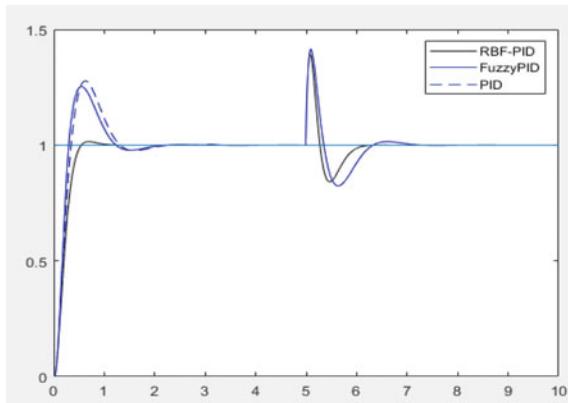
The spray control method of plant protection UAV needs fast response and smooth transition to ensure uniform spray effect, so in this section the dynamic performance of RBF-PID was tested. Simulation based on MATLAB-Simulink platform was given to validate the performance of PID, Fuzzy PID and RBF-PID. Parameters of PID are chosen as  $K_p = 1.00$ ,  $K_i = 0.20$  and  $K_d = 0.50$ . The number of nodes in input layer, hidden layer and output layer are 3, 6 and 1, respectively. The learning rate  $\eta$  is 0.20, momentum factor  $\alpha$  is 0.05, and the initial value of the weight is a random number.

#### 3.2 Analysis of Simulation Results

Take the step signal as the system input, and the response curves of the three control methods are shown in Fig. 5.

It can be seen from Fig. 5 that the settling time of RBF-PID, Fuzzy PID and PID are 0.71 s, 1.05 s and 1.15 s and the overshoot are 6.1%, 25.32% and 27.7%,

**Fig. 5** Response curves of RBF-PID, Fuzzy PID and PID



respectively. So, the response speed of RBF-PID is obviously quicker than the other two methods and the transition process is smoother. It can be inferred that RBF-PID satisfied the requirements of variable spray.

## 4 Control Accuracy Experiment and Response Speed Test

### 4.1 Control Accuracy Experiment

The flight speed of plant protection UAV is 2–6 m/s in general during spraying operation. In view of Eq. (2), it can be indicated that when the spray swath and pesticide application volume per unit area of UAV are fixed, the desired spray flow of the system is directly proportional to the flight speed of UAV. Therefore, the actual performance of the system is verified by measuring the actual flow volume of the variable spray system at different speeds. The specific test steps were as follows: at first, spray swath width of UAV was set to 5.5 m and pesticide application volume per unit area was set to 15L/hm<sup>2</sup>, then the actual spray flow of the UAV at different speeds was measured, and finally, the relative error between the desired spray flow and the actual spray flow was calculated [18]. In order to ensure the preciseness and reliability of the results, we repeated the measurement three times at the same speed, and took the average of the three measurement results as the effective value.

From Table 1, it can be seen that when the velocity of plant protection UAV is in the range of 2.0–6.0 m/s, the relative error between the actual spray flow and the desired spray flow does not exceed 10%. The result shows that the system designed in this paper can control the actual spray flow to follow the desired spray flow. Combined with the actual situation of UAV plant protection operation, the variable spray control system can adjust the flow well and basically meets the needs of current plant protection UAV spraying operation.

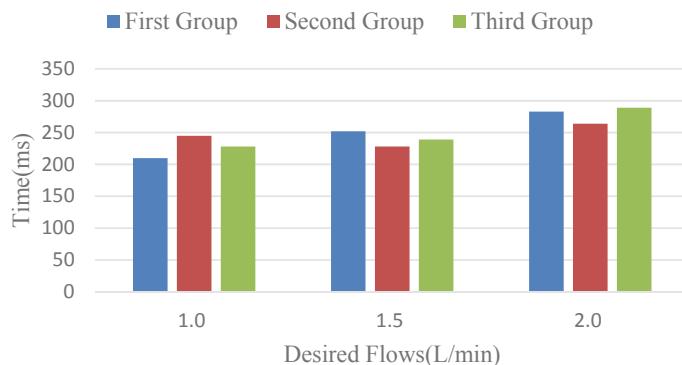
**Table 1** System flow data

Velocity (m/s)	Desired spray flow (L/min)	Actual spray flow (L/min)	Relative error (%)
2.00	1.00	1.07	7.00
2.50	1.25	1.32	5.60
3.00	1.50	1.55	3.33
3.50	1.75	1.59	9.14
4.00	2.00	1.83	8.50
4.50	2.25	2.08	7.56
5.00	2.50	2.39	4.40
5.50	2.75	2.83	2.91
6.00	3.00	2.78	7.33

## 4.2 Response Speed Test

In order to measure the response time of the system under different desired flows, different desired flow values were set in the program to test the time required to reach the set flow. The cut-off point of the collected response time is the time point when the collected flow does not change. Each test was repeated three times.

It can be seen from Fig. 6 that the response time of the system is less than 300 ms under different desired flows, so the response speed of the variable spray system designed in this paper is fast, which can meet the requirements of variable spray.

**Fig. 6** Response time of the system under different desired flows

## 5 Conclusion

In order to improve the uniformity and stability of the spray deposition, the real-time speed of UAV was obtained by combining GPS and accelerometer. On this basis, plant protection UAV variable spray system based on RBF-PID control was designed, and the RBF-PID control method can achieve stable and accurate regulation of spray flow. The simulation results show that compared with PID and fuzzy PID, the settling time of RBF-PID is shortened by 0.44 s and 0.34 s, and the overshoot is smaller. Therefore, the dynamic performance of RBF-PID is better than that of PID and fuzzy PID, which can better meet the demands of rapidity and stability for variable spray system in UAV operation. The control accuracy experiment and response speed test show that both the control precision and response speed of the system meet the requirements of variable spray.

In the process of flow regulation, this system did not consider the influence of pressure on the flow control in the whole spray process. In the future research work, the pressure will be added as parameters to realize the adjustment of spray flow.

**Acknowledgements** The authors acknowledge financial supported from the National Key Research and Development Plan(No.2020YFD1000202). Additionally, the contributions of Weihong Liu, Zhili Gong, and Changjian Yuan to this study are highly appreciated.

## References

1. Xiongkui, H.: Spray system and pesticide application technology of plant protection UAV in China. *Agricul. Eng. Technol.* **38**(09), 33–38 (2018)
2. Baijing, Q., Run, Y., et al.: Research progress of variable rate spray technology. *Trans . Chinese Soc. Agricult. Mach.* **46**(3), 59–72 (2015)
3. Sheng, W., Quanyong, Z., et al.: Design of plant protection UAV variable spray system based on neural networks. *Sensors (Basel, Switzerland)* (2019)
4. Yanggang, L., Yu, R., et al.: Model and design of real-time control system for aerial variable spray. *PLoS One* **15**(7), e023570 0(2020)
5. Yanlei, X., Jialin, B., et al.: Design and experiment of multi nozzle combined variable spraying system. *Trans. Chinese Soc. Agricult. Eng.* **32**(17), 47–54 (2016)
6. Yubin, L., Guobin, W.: General situation and prospect of UAV industry in China. *Agricult. Eng. Technol.* **38**(09), 17–27 (2018)
7. Zhenzhao, C., Xuejun, Y., et al.: Design and experiment of adaptive variable spray system for UAV Based on neural network PID. *J. South China Agricult. Univ.* **40**(04), 100–108 (2019)
8. Jing, P.: Research on UAV positioning technology based on multi sensing information fusion. *South China Univ. Technol.* (2020)
9. Lebeau, F., Destain, B., et al.: Improvement of spray deposit homogeneity using a PWM spray controller to compensate horizontal boom speed variations. *Comput. Electron. Agricult.* **43**(2), 149–161 (2004)
10. Sheng, W., Quanyong, Z., et al.: Design and experiment of a variable spray system for unmanned aerial vehicles based on PID and PWM control. *Appl. Sci.* **8**(12), 2482 (2018)
11. Fritz, B.K., Hoffmann, W.C., et al.: Effects of spray mixtures on droplet size under aerial application conditions and implications on drift. *Appl. Eng. Agric.* **26**(1), 21–29 (2010)

12. Wenfeng, S., Haiyang, L., et al.: Design and experiment of PID control variable spraying system based on neural network tuning. *Trans. Chinese Soc. Agricult. Mach.* **51**(12), 55–64, 94 (2020)
13. Dashuai, W., Junxiong, Z., et al.: Design and experiment of plant protection UAV dynamic variable spraying system. *Trans. Chinese Soc. Agricult. Mach.* **048**(005), 86–93 (2017)
14. Yannong, L., Tinglan, L., et al.: Adaptive PID control of quadrotor aircraft based on RBF neural network. *Control Eng. China* **23**(3), 378–382 (2016)
15. Songwei, Z., Haigen, H., et al.: Nonlinear adaptive PID control for greenhouse environment based on RBF network. *Sensors* **12**(5), 5328–5348 (2012)
16. Mingguang, Z., Xinggui, W., et al.: Adaptive PID control based on RBF neural network identification. In: IEEE International Conference on Tools with Artificial Intelligence (2005)
17. Liping, H.: Study on electrokinetic variable rate fertilization control system based on genetic optimization RBF-PID. Heilongjiang Bayi Agricultural University (2020)
18. Ruirui, Z., Yang, L., et al.: Design and experiment of variable rate spraying control system for manned helicopter. *J. Agricult. Mech. Res.* **39**(010), 124–127 (2017)

# Research on a Safe and Reliable Agricultural Product Traceability System Driven by Permissioned BlockChain Technology



Guofeng Zhang , Xiao Chen , Bin Feng , and Juan Wen

**Abstract** This article analyzes and summarizes the current 5W1H problems faced by the Agricultural Product Traceability System. In order to solve these problems, based on the Permissioned BlockChain technology and Cryptographic technology, a logical framework for the safety and reliable Agricultural Product Traceability System is designed. The framework uses data encryption and flexible access control to protect data privacy and security of participants in the traceability system. A safe and reliable three-dimensional agricultural product traceability mechanism of “Combination of Peacetime and Wartime” + “Criss-Cross” is proposed. With the help of Smart Contracts, the agricultural products traceability data can be shared safely in time and matter. Effectively solve the problem of not being able to obtain real traceability data. It can not only trace the responsible party, but also ascertain the truth of the security incident. The research results of this paper provide theoretical support for the research of Agricultural Product Traceability System.

**Keywords** Agricultural product traceability system · Permissioned BlockChain · Safety · Reliable

## 1 Introduction

In recent years, food safety incidents occur frequently all over the world, and countries attach great importance to the quality and safety assurance system of agricultural products and food. If consumers want to eat safely and healthily, they need

---

G. Zhang ( ) · B. Feng

School of Information Science and Technology, Taishan University, Taian 271000, Shandong, China

e-mail: [zhangguofeng@tsu.edu.cn](mailto:zhangguofeng@tsu.edu.cn)

X. Chen

School of Economics and Management, Taishan University, Taian 271000, Shandong, China

J. Wen

College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

to ensure the quality and safety of agricultural products. The safety of agricultural products is an important factor to maintain people's life and health. The establishment of a safe and reliable Agricultural Product Traceability System (APTS) is an important prerequisite to ensure people's health. Therefore, the government or agricultural regulatory agencies need to accelerate the construction of the quality and safety traceability system of agricultural products and strengthen the supervision of the quality and safety of agricultural products [1]. A recent IBM study shows that 71% of consumers say traceability is important to them and are willing to pay a premium for brands that provide traceability. BlockChain technology relies on its non-tamperable and open and transparent technical characteristics to build a food safety traceability platform, which can enhance product information transparency and consumer confidence, gain consumer trust, and further improve the food safety governance ecosystem [2]. In the traceability application scenario, the application of BlockChain technology in food safety governance is the most common and typical [2]. Permissioned BlockChain is a kind of BlockChain that each node needs to be licensed by regulatory agencies or authoritative organizations, especially suitable for agricultural products traceability and supervision.

From the perspectives of consumers, government regulatory agencies, and participants in the agricultural product supply chain, this article explores the logical architecture, data privacy protection, data sharing and supervision of a safe and credible APTS. The goal is to provide consumers with safe and reliable agricultural product traceability services, provide flexible, efficient and accurate regulatory services for government regulatory agencies, and protect the data security and privacy of participants in the agricultural product supply chain. Finally, build a APTS with active participation of multiple users, data security and reliable, and efficient and flexible supervision.

## 2 Relate Works

There are many construction plans for the APTS, but they are inseparable from the legal basis, traceability technology and traceability objects of agricultural product traceability. This paper summarizes and analyzes the research and application status of APTS from the three dimensions of agricultural product traceability laws and regulations, technology, and category.

### 2.1 Laws and Regulations

From a global perspective, the European Union, the United States, Japan, and China have successively issued a series of laws and regulations to regulate and strengthen the supervision of the quality and safety of agricultural products. Japan began to promote the food traceability system in 2001, and began to trace the entire supply chain of beef

in 2002. The European Union promulgated the general food law No. 178/2002 in 2002, stipulating that from 2004 onwards, within the EU All foods sold can be tracked and traced. The United States began to implement “farm-to-table” risk management in 2002, requiring companies to establish a traceability system. China has also issued a number of laws and regulations. In 2015, it required National important products such as agricultural products and food can be traced [3]. Obviously, the quality and safety of agricultural products is a universal problem worldwide, which must be regulated and guided through legislation to ensure consumer food safety.

## 2.2 *Tracing Technologies*

Generally speaking, the circulation of commodities needs to go through multiple links and multiple subjects. The research believes that at least five circulation systems from the producer to the retailer can be delivered to the final consumer [4]. Agricultural product traceability is a complex process that runs through production, processing, warehousing, logistics, and sales. The entire process information from planting/breeding, processing, logistics, and sales is a strong guarantee for food safety [4], and information processing must be done with the help of multiple technologies. The current mainstream traceability technologies include barcode, RFID, geographic information system, WEB service and other technologies [3]. With the continuous development and application of Internet technology, the current quality and safety traceability of agricultural products mainly depends on the realization of QR codes or barcodes.

The traditional traceability system adopts a centralized architecture, and data is stored in a database centrally. There are problems such as data loss, privacy leakage, and insufficient authority. BlockChain has technical features such as distributed storage, non-tamperable data, and traceability. If the BlockChain is used effectively, the food supply chain can become more transparent and traceable. At present, BlockChain technology has also been researched and applied in the traceability system. For example, Salah proposed a BlockChain-based solution and framework [5]. The Sku-chain research uses BlockChain technology to track the entire transportation process of agricultural products [6]. Based on Web technology, built a traceability system of IPFS + Ethereum [3].

Looking to the future, from the perspective of traceability technology, BlockChain will be deeply integrated with technologies such as the Internet of Things, Big Data, and Artificial Intelligence to build an efficient, safe and reliable traceability system.

## 2.3 *Agricultural Products Traceability Category*

1. Agricultural products in plantation industry. Huawei regards the agricultural BlockChain as an important part of Huawei’s “Agricultural Fertile Soil Cloud

- Platform”, opening up multiple links from sowing to production and processing, to circulation and sales, and building a “seed to table” agricultural product traceability system. Salah realizes traceability and visibility in the soybean supply chain based on BlockChain technology [5]. Tao et al. [7] applied BlockChain technology to ensure the food safety of grain rice, and provided a tracking system for traceability information from farmland to table. Gao et al. established a tea quality and safety traceability system based on BlockChain [3].
2. Livestock and poultry breeding. Alibaba Cloud and ZhongAn Technology Company’s BlockChain farming- “Bubu Chicken” project automatically collects the location of the chicken and uploads it to the BlockChain in real time. Traceability information includes chicken breeds, environmental sensor data, weight, health, growth cycle, slaughter data, quarantine, sales information, etc., all of which are stored in the BlockChain. The quality of broiler products of Shuguang Farm can be traced, and the traceability can be traced to the house breeding and traceability to the primary distributors and consumers, realizing the whole process of monitoring from breeding, processing to consumer terminals.
  3. Fresh/cold chain food. Wal-Mart and IBM designed a food traceability system based on Hyperledger Fabric to ensure consumers’ confidence in food safety. It adopted a distributed food supply ecosystem and applied BlockChain technology to logistics and supply chain management [8]. Yao et al. [4] designed a cold chain food traceability system based on the Fisco Consortium BlockChain. Zhao et al. [9] built a framework model of a fresh food mobile traceability platform based on the dual-chain architecture of account BlockChain and transaction BlockChain. This shows that BlockChain technology can be used for cold chain food traceability management, and can track and solve the problems detected in the subsequent circulation process in time [4].
  4. Fishery agricultural products. The Norwegian Seafood Association cooperated with IBM and Atea to create a salmon traceability system based on BlockChain. The camera tracks the salmon’s environmental information and logistics information during its life cycle and stores it in the BlockChain, so that customers can clearly know which fjord the fish came from, when they were caught, as well as the feed the fish eat and whether the facility is Use sustainable methods. Jun et al. [10] took fish and meat products as an example to explore the feasibility of the practical application of a BlockChain-based agricultural product traceability system.

From the perspective of agricultural product traceability categories, it has basically covered many, but the existing schemes are all at the traceability level of agricultural products. Since the post-processing of agricultural products may involve multiple categories, the traceability of agricultural products in the future will move from the main agricultural products to the full category coverage.

### 3 Discussion on the Traceability of Agricultural Product

#### 3.1 Centralized System Architecture Issues

The existing traceability system research is mainly based on a centralized system architecture, which is self-built by core enterprises or authoritative organizations, and the data is centrally stored and independently owned by the system builder. From the perspective of traceability business, driven by interests, system owners are likely to tamper with data at will in order to evade responsibility for the incident, resulting in the inability to correctly trace the source of the incident. At the same time, there are external risks. The traceable data is easily stolen, and the information is used as shoddy, resulting in the proliferation of low-quality and high-hazard food in the consumer market. From the perspective of technical implementation, once a hacker attack or an accident such as storage media damage occurs, it will cause a single point of failure and fail to provide effective services [11]. The most important thing is that when a food safety incident occurs, the centralized system architecture can only provide the traceability data and results of a single subject. Its authenticity needs to be verified, which makes it impossible to restore the truth of the incident, thus losing the authority and credibility of the traceability system.

#### 3.2 Distributed Architecture Based on BlockChain

It has become possible to use the distributed storage of the BlockChain to solve the above problems, thereby building a BlockChain-based agricultural product traceability system, and providing technical endorsement for agricultural product quality and safety trust issues [3]. For example, through BlockChain technology, a decentralized agricultural product traceability system can be established to improve the traceability of agricultural products and strengthen the security and transparency of the agricultural supply chain. However, if the data stored on the BlockChain is open and transparent, when the information on the production, circulation, and transaction of agricultural products remains on the chain, the privacy and commercial secrets of the data owner will be leaked. For example, a competitor maliciously analyzes the variety, output, and listing date of agricultural products. Once the competitor is fully grasped, it will damage the interests of producers by controlling partial prices, resulting in increased production but no increase in revenue [1]. At the same time, due to the limited block size and processing speed of the BlockChain, if all data is stored on the chain, the overall performance of the system will be reduced. Therefore, choose which agricultural product data to upload in real time and which data to upload when needed. How to protect the privacy and security of data stored on the chain and retain control rights, and to facilitate the flexible sharing of data on the chain. These are the problems to be solved by the distributed agricultural product traceability system based on the BlockChain.

### 3.3 5W1H Problem Model

Although some progress has been made in the research of agricultural product traceability systems, the centralized architecture and BlockChain-based distributed architecture still face many problems. From the perspective of application promotion effects, individuals in the agricultural product supply chain are relatively independent, information transmission is blocked, and a complete security and trustworthy system and privacy protection system have not yet been established. Through the above analysis, this article summarized the 5W1H agricultural product traceability system problem model, as shown in Table 1.

The above analysis table shows that the problems faced by agricultural product traceability include both technical problems and the participants' own problems. However, the root cause of the problems lies in two points:

1. The existing technical solutions for the traceability system of agricultural products cannot guarantee the privacy and commercial secrets of the participants in

**Table 1** The 5W1H problem model of APTS

Definition	Question	Reason
Who	Who has the problem?	Participants: fear of divulging business secrets, afraid of data sharing and data fraud etc.
		Regulatory: lack of traceability mechanism and means
		Production/consumer: weak legal awareness and food safety awareness
When	When to trace agricultural products?	Peacetime: normal purchase and consumption Wartime: the outbreak of agricultural product quality and safety incidents
Why	Why trace agricultural products?	Peacetime: confirm the quality and safety of agricultural products Wartime: quick recall of agricultural products; clarify the responsible party; find out the root cause of the problem
Where	Where is the difficulty of the problem?	Lack of real and available data Compatibility of data confidentiality and flexible authorization sharing
What	What is the root of the problem?	The participants are untrustworthy (malicious competition, disclosure of trade secrets, shirking responsibility). Untrustworthy between consumers and participants (dominated by strong side)
How	How is the traceability effect?	It is easy to know the circulation path, but it is difficult to find the root cause of the problem

the supply chain, resulting in the inability of mutual trust between participants and the inability to actively participate in the construction of the traceability system.

2. The credible traceability mechanism of agricultural products is not complete, the plan is incomplete, and it is impossible to provide consumers with safe and credible traceability services, and it is impossible to trace the source when a food safety incident occurs, that is, between the consumer and the agricultural product traceability system There is no guarantee of safety and mutual trust.

To realize the safety and usability of the APTS, the above two problems must be solved. When the participants of the traceability system trust each other in security, the participants are willing to share their real data publicly. When consumers believe that the traceability results are safe and reliable, the consumers will accept the traceability results of agricultural products from their hearts and actively participate in the construction of the traceability system. Further promote the healthy development of the agricultural product traceability system.

## 4 Design of Safe and Reliable Agricultural Products Traceability System

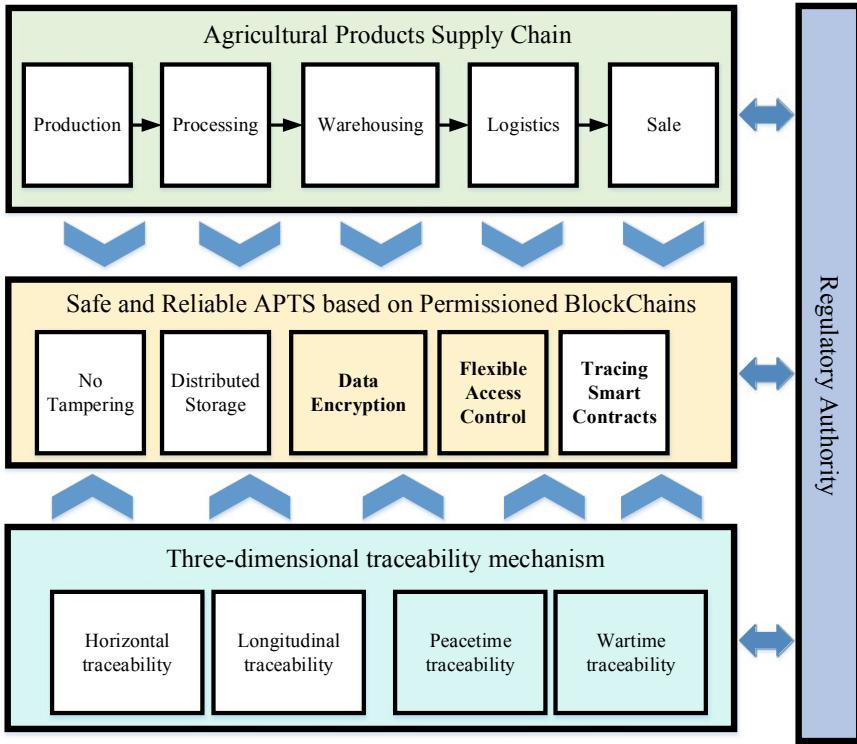
### 4.1 Safe and Reliable Logical Framework

Based on the technical advantages of decentralization, tamper-proofing, openness and transparency of BlockChain technology, various entities in the agricultural product supply chain can achieve secure interactive sharing of information through pre-set rules, smart contracts, and information sharing technologies [1]. Based on the identity authentication and authorization of Permission BlockChain technology, regulatory agencies can effectively and accurately supervise system participants. Based on Permission BlockChain technology and cryptography technology, this paper designs a safe and credible agricultural product traceability system logical architecture, as shown in Fig. 1.

A safe and reliable APTS covers the entire process of production, processing, warehousing, logistics, and sales in the agricultural product supply chain (APSC). All participants in the APSC can carry out normal business under the effective supervision of the regulatory authority. The regulatory agency is responsible for the identity authentication, authority management, data supervision, and traceability of agricultural product quality and safety events for each participant.

The core features of this framework are in two aspects:

First, adopt the CP-ABE encryption scheme to protect data privacy and realize safe sharing. It is to encrypt and store agricultural product traceability data, and at the same time, realize the authorized access and sharing of encrypted data with the help of flexible access control technology. Using the Attribute-based Encryption

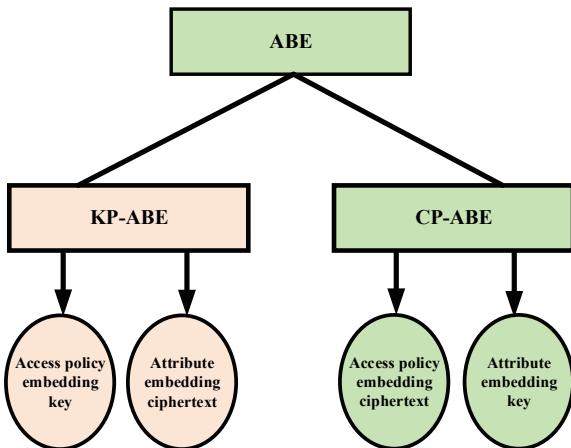


**Fig. 1** Safe and reliable logical framework of APTS

(ABE) scheme proposed by Sahai and Waters [12], a series of attribute sets can be used instead of unique identifiers to identify identities, and access control rules and encryption can be configured with the help of attributes. The main thing is to realize the binding of encrypted data and access control. Only users who have the corresponding attributes of encrypted data and meet their access control conditions can decrypt the data, which can fully satisfy the 1-to-N fine-grained access control of encrypted data.

As shown in Fig. 2, the ABE encryption scheme is further divided into two types, KP-ABE [13] and CP-ABE [14], according to specific implementations. Among them, the CP-ABE scheme can embed the access control strategy into the ciphertext, and the attributes are used as the access control strategy and the key. It is more suitable for encrypted data sharing scenarios. For example, the CP-ABE scheme has been used in the EHR fine-grained traceability scheme [15]. Therefore, the CP-ABE scheme is used in the scheme of this article for data encryption and access control, protects its data privacy, and guarantees access control rights to shared data. Therefore, the data security concerns of the data issuer are eliminated, and the security and credibility between participants are realized.

**Fig. 2** Classification of ABE encryption schemes



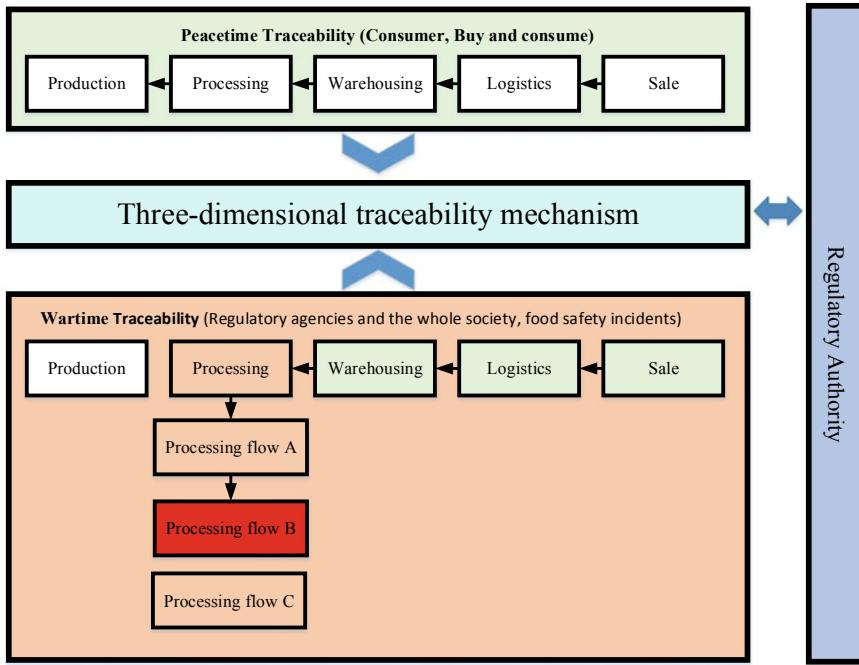
Second, the traceability Smart Contract is a concrete realization of the agricultural product safety and reliable traceability mechanism, which can realize the traceability of agricultural products in different scenarios. When a food safety incident breaks out, offline data of relevant participants can be automatically uploaded to the chain in real time. On the one hand, it can prevent the party responsible for the event from forging or tampering with real business data in an emergency. On the other hand, all complete transaction data can be stored off-chain, but the data summary and storage address need to be uploaded to the chain in real time to ensure the authenticity of the data.

## 4.2 Safe and Reliable Three-Dimensional Traceability Mechanism

The traceability mechanism is the core infrastructure that guarantees the availability and credibility of the agricultural product traceability system. In order to solve the problems faced by the agricultural product traceability system, this paper studies the establishment of a three-dimensional agricultural product credible traceability mechanism of “Combination of Peacetime and Wartime” + “Criss-Cross. The process is shown in Fig. 3.

The above-mentioned traceability mechanism includes two types: Peacetime traceability and Wartime traceability. The specific functions are as follows:

Peacetime traceability is for ordinary consumers. When purchasing and consuming, it can be traced back to the circulation process of agricultural products, that is, horizontal traceability, to check whether each process is safe to ensure consumers' right to know. In this scenario, consumers only pay attention to whether the production process of agricultural products meets the standards and whether they



**Fig. 3** Three-dimensional traceability mechanism

are authentic brands or quality. Relatively little attention is paid to the process data in the production process of agricultural products, such as the light, temperature, water, gas and fertilizer in the growth of crops, and the pH in the processing process. This information requires professionals to determine whether it meets the production standards. It doesn't make much sense. However, when a food safety incident breaks out, this information is sufficient to determine the responsibility of the relevant participants. Through further detailed analysis, the root cause of the problem can be analyzed. However, this information often belongs to the core data and commercial secrets of participants in various links, and cannot be made public unless forced to do so.

In wartime traceability, when a food safety incident breaks out, regulatory agencies and the whole society will keep an eye on it. Although producers are unwilling to disclose the core data, in order to find out the cause of the incident, the complete data of the incident must be disclosed to the society. Therefore, the regulatory agency must disclose the true and complete data to the society as soon as possible. Under the architecture based on Permissioned BlockChain technology, with the help of Smart Contract, there is no need for human attention and automatic upload, which solves the problem that the traditional traceability system cannot obtain complete data in the first time. As shown in Fig. 3, when a food safety incident broke out, a horizontal tracing process first found that there was no problem from the sales link to

the warehousing link. The problem occurred in the processing link. Then, according to the processing data, the problem in the processing process B was obtained. So as to realize the vertical traceability.

The above-mentioned three-dimensional credible traceability mechanism of agricultural products completely solves the problem of not only investigating the responsibility, but also the root cause of the problem in the traceability of agricultural products. At the same time, it realizes the on-demand sharing of business data in time-sharing and division of tasks, ensuring user data privacy, and improving the security and credibility of the system.

## 5 Conclusions

Based on the analysis of the research status of APTS, this paper summarizes and refines APTS's 5W1H problem model. In order to eliminate APTS participants' worries about the privacy and security of shared data, from the perspective of protecting data privacy and security, a safe and reliable APTS framework and a three-dimensional safe and reliable agricultural product traceability mechanism have been proposed. The above research results are helpful to solve the problems such as the participants of the supply chain dare not share data, illegally tamper with data, security incidents cannot "get to the bottom". They are an effective supplement and important promotion to the existing BlockChain-based APTS research, and provide theoretical and policy support for the privacy protection and secure sharing of agricultural product traceability data.

**Acknowledgements** This work was supported by the Project of National Natural Science Foundation of China (Grant no. 62071320, Grant no. 61771090), Shandong federation of social sciences (Grant no. 2021-YYGL-32) and Tai'an Science and Technology Innovation Development Project (Grant no. 2020NS080).

## References

1. Chen, C., Lin, W., Chuanbo, L.: Research on optimization path of safety and economic benefits of agricultural products in Yangzhou under the mode of "blockchain+agriculture. Market Wkly. **34**(04), 1–3 (2021)
2. Ruohong, Z., Jun, L., Mingde, X.: The innovative applications of BlockChain technology in market regulation and its challenges. Manage. Technol. SME **01**, 194–196 (2021)
3. Qijuan, G., Chunjie, Y., Xianchun, W.: Research on the traceability system of tea quality and safety based on blockchain. J. Anhui Agric. Univ. **48**(2), 299–303 (2021)
4. Chao, Y., Song, T.: Research on the application of BlockChain technology in cold chain traceability. J. Hebei Acad. Sci. **38**(01), 78–83 (2021)
5. Salah, K., Nizamuddin, N., Jayaraman, R., et al.: BlockChain-based soybean traceability in agricultural supply chain. IEEE Access **7**, 73295–73305 (2019)

6. Allison, L.: Skuchain: Here's how BlockChain will Save Global Trade a Trillion Dollars. International Business Times, 2016 [EB/OL]. <https://www.skuchain.com/skuchain-heres-how-BlockChain-will-save-global-trade-a-trillion-dollars/>
7. Qi, T., Xiaohui, C., Siming, Z., et al.: The food quality safety management system based on BlockChain technology and application in rice traceability. *J. Chin. Cereals Oils Assoc.* **33**(12), 102–110 (2018)
8. Iftekhar, A., Cui, X., Hassan, M., Afzal, W.: Application of blockchain and internet of things to ensure tamper-proof data availability for food safety. *J. Food Qual.* (2020)
9. Lei, Z., Xinhua, B., Anni, Z.: Frame reconstruction of mobile traceability information system for fresh foods based on blockchain. *Food Sci.* **41**(3), 314–321 (2020)
10. Jun, S., Xiaodong, H., Jianhua, C.: Research on the agricultural products safety traceability based on the BlockChain. *J. Henan Agric. Sci.* **47**(10), 149–153 (2018)
11. Yanjun, T., Longyi, A., WenHong, X.: Construction of edible fungus supply chain digital platform based on blockchain technology. *Heilongjiang Agric. Sci.* **4**, 111–114 (2021)
12. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: International Conference on Theory & Applications of Cryptographic Techniques, pp. 457–473 (2005)
13. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, pp. 89–98 (2006)
14. Bethencourt, J., Sahai, A., Waters, B.: Ciphertext-policy attribute-based encryption. In: Proceedings Security and Privacy, pp. 321–334 (2007)
15. Zuobin, Y., Yuaping, S., Jianfeng, M., et al.: Blockchain-based distributed EHR fine-grained traceability scheme. *J. Commun.* 1–11 (2021)

# Simultaneously Learning Syntactic Dependency and Semantics Reasonability for Relation Extraction



Xin Wang , Nan Yin , Xiang Zhang , Xinyi Bai , and Zhigang Luo

**Abstract** Relation extraction as an important Natural Language Processing (NLP) task is to identify relations between named entities in text. Recently, graph convolutional networks over dependency trees have been widely used to capture syntactic features and achieved attractive performance. However, most existing dependency-based approaches ignore the positive influence of the words outside the dependency trees, sometimes conveying rich and useful information on relation extraction. In this paper, we propose a novel model, Entity-aware Self-attention Contextualized GCN (ESC-GCN), which efficiently incorporates syntactic structure of input sentences and semantic context of sequences. To be specific, relative position self-attention obtains the overall semantic pairwise correlation related to word position, and contextualized graph convolutional networks capture rich intra-sentence dependencies between words by adequately pruning operations. In this way, our proposed model not only reduces the noisy impact from dependency trees but also obtains easily-ignored entity-related semantic representation. Extensive experiments demonstrate that our model achieves encouraging performance.

**Keywords** Relation extraction · Self-attention · Dependency trees · Semantic representation

---

X. Wang · N. Yin · X. Zhang · X. Bai · Z. Luo

College of Computer, National University of Defense Technology, Changsha, China  
e-mail: [wangxin19@nudt.edu.cn](mailto:wangxin19@nudt.edu.cn)

X. Zhang

Institute for Quantum and State Key Laboratory of High Performance, Computing National University of Defense Technology, Changsha, China

Z. Luo

Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha, China

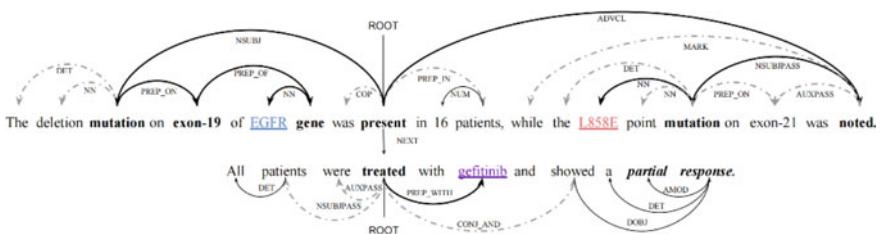
## 1 Introduction

There has been major interest in relation extraction, which aims to assign a relation among a pair of entity mentions from plain text. Recent models for relation extraction are primarily built on deep neural networks, which encode the entire sentence to obtain relation representations and have made great progress [1, 2].

From the sample given in Fig. 1, there is a relation “sensitivity” between the three entities within the two sentences, which expresses that “tumors with L858E mutation in EGFR gene respond to gefitinib treatment”. Prior efforts show that models utilizing dependency parsing of input sentences are very effective in relation extraction because their superiority lies in drawing direct connections between distant syntactically correlated words. Xu et al. [3] first applied LSTM on the shortest dependency path (SDP) between the entities in the full tree. Miwa et al. [4] reduced the full tree into the subtree below the lowest common ancestor (LCA) of the entities. Both patterns prune the dependency trees between the entities to cover relevant information and discard noises. However, if only consider the dependency structure shown in Fig. 1 (i.e., SDP, LCA), the token “*partial response*” will be neglected, yet, they contribute to the gold relation greatly. Therefore, it is very essential to obtain the interactions of all words, not just the dependency trees of entities. To address this issue, we use a relative position self-attention mechanism, which allows each token to take its left and right context into account while calculating pairwise interaction scores with other tokens.

Recently, combining entity position features with neural networks has greatly improved the performance of relation extraction. Zhang et al. [5] combined sequence LSTM model with a position-attention mechanism and got a competitive result. From their experiments, we know that the words determining relation are frequently related to the target entities. However, these methods only utilize the semantic representations and position features, ignoring the dependency syntax of the words. Unlike previous efforts, which focus on either dependency parsing or the semantic features, we synthesize syntactic dependency structure and entity-related sequential semantic context into an attention mechanism, both of which are crucial for relation extraction.

In this paper, we first utilize relative position self-attention mechanism to encode semantic interactions of the sequential sentence, which ignores the distance between words to calculate the compatibility scores and relative position scores. Then



**Fig. 1** Example of dependency parsing for two sentences expressing an interaction

contextualized graph convolution module encodes the dependency trees between the entities to capture contextual long-range dependencies of words. Afterwards, entity-aware attention mechanism combines these two modules to get final relation representations. The contributions of our work are summarized as follows:

1. We propose a ESC-GCN model to learn relation representations. Compared with previous methods, our method not only utilizes semantic features but also considers dependency features.
2. Our proposed model proves to be very competitive on the sentence-level task (i.e., TACRED and SemEval dataset) and cross-sentence n-ary task. Especially, our model outperforms most baseline in long sentences.
3. We show that our model is interpretable by visualizing the relative position self-attention.

## 2 Related Work

Relation extraction has been intensively studied for a long history, and most existing neural relation extraction models can be divided into two categories: sequence-based and dependency-based. Zeng et al. [1] first applied CNN with manual features to encode relations. Zhang et al. [2] first applied RNN to relation extraction and got competitive performance. Zhang et al. [6] employed BiLSTM to learn long-term dependencies between entity pairs. However, these models only considered the sequential representations of sentences and ignored the syntactic structure. In fact, These two characteristics complement each other actually.

Compared with the sequence-based models, incorporating dependency syntax into neural models has proven to be more successful, which captures non-local syntactic relations that are only implicit in the surface from alone [7, 8]. Xu et al. [3] purposed SDP-LSTM that leverages the shortest dependency path between two entities. Peng et al. [9] proposed a graph-structured LSTM for cross-sentence n-ary relation extraction, which applied two directed acyclic graphs (DAGs) LSTM to capture inter-dependencies in multiple sentences. Song et al. [10] proposed a graph-state LSTM model which employed a parallel state to model each word, enriching state scores via message passing. Zhang et al. [11] presented C-GCN for relation extraction, which uses graph convolution and a path-centric pruning strategy to selectively include relative information.

In recent past, Vaswani et al. [12] proposed an attention model called Transformer. Verga et al. [13] used self-attention to encode long contexts spanning multiple sentences for biological relation extraction. Zhang et al. [5] employed a position attention mechanism over LSTM outputs for improving relation extraction. Bilan et al. [14] substituted the LSTM layer with the self-attention encoder for relation extraction. Yu et al. [15] proposed a novel segment attention layer for relation extraction and achieved competitive results on TACRED dataset.

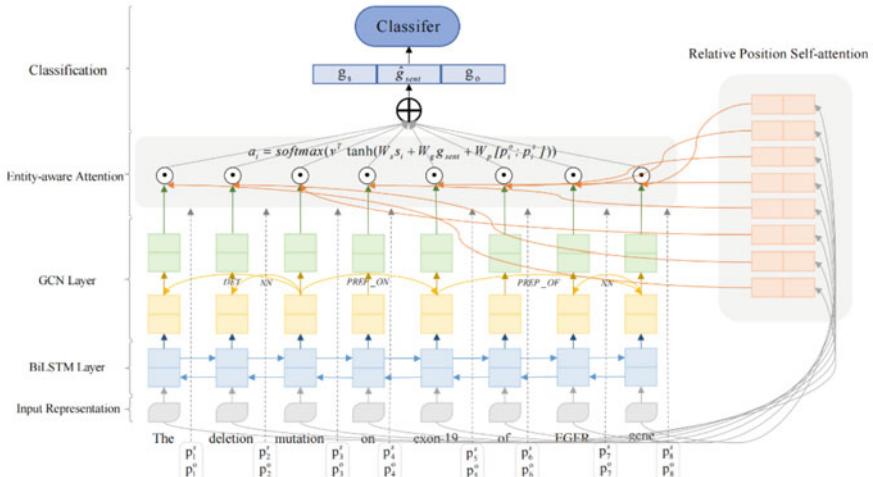
### 3 The Proposed Model

We define relation extraction as a multi-class classification problem, which can be formalized as follows: Let  $S = [w_1, w_2, \dots, w_n]$  denote a sentence, where  $w_i$  is the  $i$ -th token. A subject-entity and an object-entity are identified:  $W_s = [w_{s_1}, w_{s_2}, \dots, w_{s_n}]$  and  $W_o = [w_{o_1}, w_{o_2}, \dots, w_{o_n}]$ . Given  $S$ ,  $W_s$  and  $W_o$ , the objective of relation extraction is to predict a relation  $r \in R$  or “no relation”. Specifically, Fig. 2 indicates the overall architecture of our work.

#### 3.1 Preliminary

GCNs are neural networks that operate directly on graph structures, which are an adaptation of convolutional networks. Given a graph with  $n$  nodes, we generate the graph with an  $n \times n$  adjacency matrix  $A$  where  $A_{ij} = 1$  if there is an edge going from node  $i$  to node  $j$ , otherwise  $A_{ij} = 0$ . We extend GCNs for encoding dependency trees by incorporating opposite of edges into the model. Each GCN layer takes the node embedding from the previous layer  $g_j^{(l-1)}$  and the adjacency matrix  $A_{ij}$  as input, and outputs updated node representation for node  $i$  at the  $l$ -th layer. Formally, the induced representation  $g_i^{(l)}$  can be defined as follow:

$$g_i^{(l)} = \rho \left( \sum_{j=1}^n A_{ij} W^{(l)} g_j^{(l-1)} + b^{(l)} \right) \quad (1)$$



**Fig. 2** Overall architecture of our proposed ESC-GCN model

where  $W^{(l)}$  is a linear transformation,  $b^{(l)}$  is the bias term, and  $\rho$  is an activation function (e.g., RELU).

### 3.2 Input Representation

In our model, the input representation module first transforms each input token  $w_i$  into a comprehensive embedding vector  $x_i$  by concatenating its word embedding  $\text{word}_i$ , entity type embedding  $\text{ner}_i$  and part-of-speech (POS) tagging embedding  $\text{pos}_i$ . Embedding vector  $x_i$  formally defined as:

$$x_i = [\text{word}_i; \text{ner}_i; \text{pos}_i] \quad (2)$$

It has already proved the words close to the target entities are generally more informative [14], we modify the position representation originally proposed by [5], and transform it into binary position encoding. Consequently, we define a binary position sequence  $[p_1^s, \dots, p_n^s]$  that relative to the subject-entity:

$$p_i^s = \begin{cases} -\log_2(s_1 - i) - 1, & i < s_1 \\ 0, & s_1 \leq i \leq s_2 \\ \log_2(i - s_2) + 1, & i > s_2 \end{cases} \quad (3)$$

Here  $s_1, s_2$  represent the start index and end index of the subject entity respectively,  $p_i^s \in \mathbb{Z}$  can be viewed as the relative distance of token  $x_i$  to the subject entity.

Similarly, we also obtain a position sequence  $[p_1^o, \dots, p_n^o]$  relative to the object entity.

### 3.3 Relative Position Self-attention Mechanism

The self-attention mechanism was firstly proposed by Vaswani et al. [12], which allows words to take its context into account. Following Bilan et al. [14], we apply several modifications to the original self-attention layer. Firstly, we simplify the residual connection that directly goes from the self-attention block to the normalization layer. Then we substitute the layer normalization with batch normalization. In our experiments, we have observed improvements with these settings, and a more detailed overview of the results can be seen in the subsection 5.1.

Traditionally, a self-attention layer takes a word representation at position  $i$  as the query (a matrix  $Q$  holds the queries for position  $i$ ) and computes a compatibility score with representations at all other positions (represented by a matrix  $V$ ). The score w.r.t. position  $i$  is reformed to an attention distribution over the entire sentence, which is used as a weighted average of representations  $E$  at all positions.

Shaw et al. [16] exploited the relative positional encoding to improve the performance of self-attention. Similarly, we modify our self-attention layer, together with a position attention that takes into account positions of the query and the object in the sentence. Our self-attention head  $s_i^{(a)}$  obtain its representation by summing pairwise interaction scores and relative position scores together, formally defined as follows:

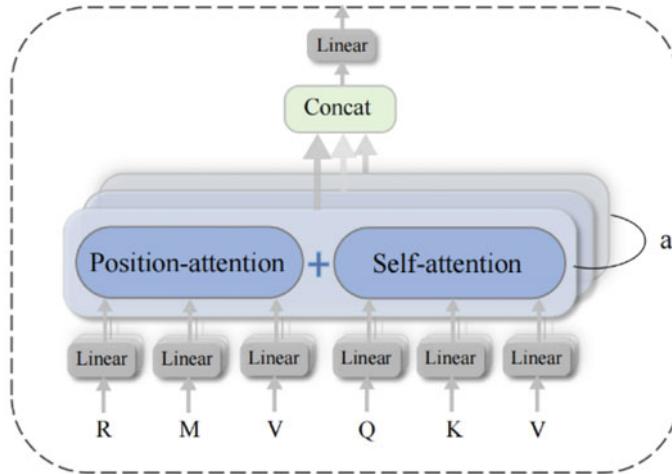
$$s_i^{(a)} = \text{softmax} \left( \frac{QK^T + RM^T}{\sqrt{d_w}} \right) V \quad (4)$$

where  $Q = W^{q(a)}e_i$ ,  $K = W^{k(a)}E$ ,  $V = W^{v(a)}E$  wherein  $W^{q(a)}, W^{k(a)}, W^{v(a)}$  are linear transformations, which map the input representation into lower dimensional space.  $d_w$  is dimension of key/query vectors which is a scaling factor same as in Vaswani et al. [12].  $M$  is relative position embedding matrix:

$$M_i = [m_{1-i}, \dots, m_{-1}, m_0, m_1, \dots, m_{n-i}] \quad (5)$$

where  $n$  is the length of the input sentence and the matrix  $M_i$  is the relative position vectors,  $m_0$  is at position  $i$  and other  $m_j$  are ordered relative to position  $i$ .

Similar to  $Q$ , we obtain a query vector  $R = W^{r(a)}e_i$  to obtain position relevance. The position attention scores result from the interaction of  $R$  with the relative position vectors in  $M_i$ . As shown in Fig. 3, we associate position attention scores with the pairwise interaction scores, which incorporates position features into overall dependencies of sequence.



**Fig. 3** Model structure of relative position self-attention

### 3.4 Contextualized GCN Layer

In this section, we construct a contextualized GCN model which takes the output from subsection 3.1 as input  $h^{(0)}$ . A BiLSTM layer is adopted to acquire the context of sentence for each word  $w_i$ . For explicitly, we denote the operation of LSTM unit as  $LSTM(x_i)$ . The contextualized word representations are obtained as follows:

$$h_i = [LSTM_{left}(x_i); LSTM_{right}(x_i)], \quad i \in [1, n] \quad (6)$$

where  $h_i \in R^{2 \times d_h}$  and  $d_h$  indicates the dimension of LSTM hidden state. Then we obtain hidden representations of all tokens  $h^{(L_1)}$ , which represents the input  $g^{(0)}$  for graph convolution, where  $L_1$  represents the layer number of RNN.

Dependency syntax has been recognized as a crucial source of features for relation extraction [7], and most of the information involved relation within the subtree rooted at the LCA of the entities. Before applying graph convolution operation, we do some tricks on the dependency parsing tree, which keeps the original dependency path in the LCA tree and incorporates 1-hop dependencies away from the subtree. Accordingly, we cover the most relevant content and remove irrelevant noise as much as possible.

Originally applying the graph convolution could bring about node representations with obviously different scales [11], since the degree of tokens varies a lot. To cope with the above limitations, we resolve these issues by normalizing the activations in the graph convolution, and add self-loop into each node in adjacency matrix  $A$ , modified graph convolution operation as follows:

$$g_i^{(l)} = \sigma \left( \sum_{j=1}^n \widetilde{A}_{ij} W^{(l)} g_j^{(l-1)} / d_i + b^{(l)} \right) \quad (7)$$

where  $\tilde{A} = A + I$ ,  $I$  is the  $n \times n$  identity matrix, and  $d_i = \sum_{j=1}^n \widetilde{A}_{ij}$  is the degree of token  $i$ . This operation updates the representation of node  $i$  by aggregating its neighborhood via a convolution kernel. After  $L_2$  iterations, we obtain the hidden outputs of graph convolution  $g^{(L_2)}$ , where  $L_2$  represents the layer number of GCN.

### 3.5 ESC-GCN for Relation Extraction

After applying the  $L_2$ -layer contextualized GCN model, we obtain hidden representation of each token, which is directly influenced by its neighbors (no more than  $L_2$  edges apart in the dependency trees). To make use of graph convolution for relation extraction, we first obtain a sentence representation as follows:

$$g_{sent} = f(g^{(L_2)}) = f(GCN(g^{(0)})) \quad (8)$$

where  $g^{(L_2)}$  denotes the collective hidden representation at layer  $L_2$  of the GCN, and  $f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^d$  is a max pooling function. Moreover, we also obtain a subject representation  $g_s$  as follows:

$$g_s = f(g_{s_1:s_2}^{(L_2)}) \quad (9)$$

as well as an object representation  $g_o$  respectively,  $s_1, s_2$  represent the start index and end index of the subject entity.

The final computation of the entity-aware attention utilizes the output state of GCN (i.e., a summary vector  $g_{\text{sent}}$ ), the self-attention hidden states output vector  $s_i$ , and the embeddings for the subject and object relative positional vectors  $p_i^s, p_i^o$ . For each hidden state  $s_i$ , attention weight  $\alpha_i$  is calculated using the following two equations:

$$u_i = v^\top \tanh(W_s s_i + W_g g_{\text{sent}} + W_p [p_i^s; p_i^o]) \quad (10)$$

$$\alpha_i = \frac{\exp(v^\top u_i)}{\sum_{j=1}^n \exp(v^\top u_j)} \quad (11)$$

where  $W_s$  weights are learned parameters using self-attention,  $W_g$  weights are learned parameters using contextualized GCN and  $W_p$  weights are learned using the positional encoding embeddings.

Afterwards,  $\alpha_i$  decides on how much each GCN outputs should contribute to the final sentence representation  $\tilde{g}_{\text{sent}}$  as follows:

$$\tilde{g}_{\text{sent}} = \sum_{i=1}^n \alpha_i g_i^{(L_2)} \quad (12)$$

Then the representation  $\tilde{g}_{\text{sent}}$ ,  $g_s$  and  $g_o$  are concatenated and fed into a feed-forward neural network (FFNN):

$$g_{\text{final}} = \text{FFNN}([\tilde{g}_{\text{sent}}; g_s; g_o]) \quad (13)$$

In the end, the final sentence representation  $g_{\text{final}}$  is then fed to another MLP layer followed by a softmax operation to obtain a probability distribution over relations:

$$p(r|g_{\text{final}}) = \text{softmax}(w \cdot g_{\text{final}} + b) \quad (14)$$

where  $g_{\text{final}}$  is the sentence representation, and  $r$  is the target relation,  $w$  is a linear transformation and  $b$  is a bias term. We utilize the cross entropy and the L2 regularization to define the objective function as follows:

$$J(\theta) = - \sum_{i=1}^s (y_i x_i, \theta) + \beta |\theta|^2 \quad (15)$$

where  $s$  indicates the total sentence;  $x_i$  and  $y_i$  represent the sentence and relation label of the  $i^{th}$  training example;  $\beta$  is  $L_2$  regularization hyper-parameter. The  $\theta$  is the whole network parameter, which can be learnable.

## 4 Experiments

### 4.1 Dataset

We follow the experimental settings in Zhang et al. [11] to evaluate our ESC-GCN model on the TACRED dataset [5] and Semeval-2010 Task 8 dataset [17]. TACRED contains over 106 k mention pairs collected from the TACKBP evaluations 2009–2014. It includes 41 relation types and a “*no relation*” class when no relation is held between entities. Mentions in TACRED are typed, subjects are classified into person and organization, and objects are categorized into 16 fine-grained classes (e.g., date, location, title).

The SemEval-2010 Task 8 dataset is an acknowledged benchmark for relation extraction (1/10 of TACRED). The dataset defines 9 types of relations (all relations are directional) and a class “*other*” denoted no relation. There are 10,717 annotated sentences which consist of 8000 samples for training and 2717 samples for testing.

### 4.2 Results on Sentence-Level Relation Extraction

We report the micro-averaged F1 scores for the TACRED dataset and the macro-averaged F1 scores for the SemEval-2010 task 8 dataset. We now report the results on the TACRED dataset in Table 1. we compare our model against following baselines: (1) sequence-based models, i.e., Convolutional Neural Networks (CNN-PE) [18], Position Aware LSTM (PA-LSTM) [5], Self-Attention Encoder (Self-Attn) [14], Segment Attention LSTM (SA-LSTM) [15]; (2) dependency-based models, i.e., the short dependency path LSTM (SDP-LSTM) [3], Tree-structured LSTM (Tree-LSTM) [19]; and (3) graph-based models: GCN and Contextualized GCN (C-GCN) [11], Simplifying Graph Convolutional Networks (S-GCN) [20].

As shown in Table 1, our ESC-GCN shows better performance than all baselines on the TACRED dataset, which achieves F1 of 67.1, outperforming C-GCN by 0.7 F1 points. This result shows the effectiveness of semantic representation. Our model obtains the highest precision, but gets a general recall value. The performance gap between Self-Attn and ESC-GCN shows that our model is better at incorporating dependency relations and the context of entities in the sentence-level task.

**Table 1** Micro-averaged precision (P), recall (R) and *F1* score on the TACRED dataset

Model	P	R	F1
CNN-PE [18]	68.2	55.4	61.1
PA-LSTM [5]	65.7	64.5	65.1
Self-Attn [14]	64.6	<b>68.6</b>	66.5
SA-LSTM [15]	68.1	65.7	66.9
SDP-LSTM [3]	66.3	52.7	58.7
Tree-LSTM [19]	66.0	59.2	62.4
GCN [11]	69.8	59.0	64.0
C-GCN [11]	69.9	63.3	66.4
S-GCN [20]	—	—	67.0
ESC-GCN (ours)	<b>71.4</b>	62.8	<b>67.1</b>

**Table 2** Macro-averaged *F1* score on SemEval-2010 task 8 dataset

Model	F1
CNN [1]	78.3
CR-CNN [21]	84.1
PA-LSTM [5]	82.7
BiLSTM + Attn [22]	84.0
SDP-LSTM [3]	83.7
SPTree [4]	84.4
C-GCN [11]	84.8
ESC-GCN (ours)	<b>85.2</b>

We also evaluate our model on the SemEval dataset, and the experimental results are shown in Table 2. Apart from above baselines, we compare ESC-GCN with other sequence-based models, i.e., Attention-based BiLSTM (BiLSTM + Attn) [22] and dependency-based models, i.e., tree-structured LSTM methods (SPTree) [4]. Our ESC-GCN achieves a competitive performance (*F1*-score of 85.2) on the SemEval dataset. The result shows that integrating entity features and dependency parsing can obtain better representation for relation extraction.

## 5 Analysis

### 5.1 Ablation Study

In order to study the contribution of each component, we conducted an ablation study on the TACRED. Table 3 shows the following results. Instead of default residual connections described by [12], the optimized residual connection contributes

**Table 3** An ablation study for ESC-GCN model

Model	F1
Best ESC-GCN	<b>67.1</b>
- Default residual	66.9
- Layer normalization	66.7
- Entity-aware module	66.6
- Relative position self-attention	66.4
- hs, ho, and feedforward (FF)	66.2
- BiLSTM layer	64.8

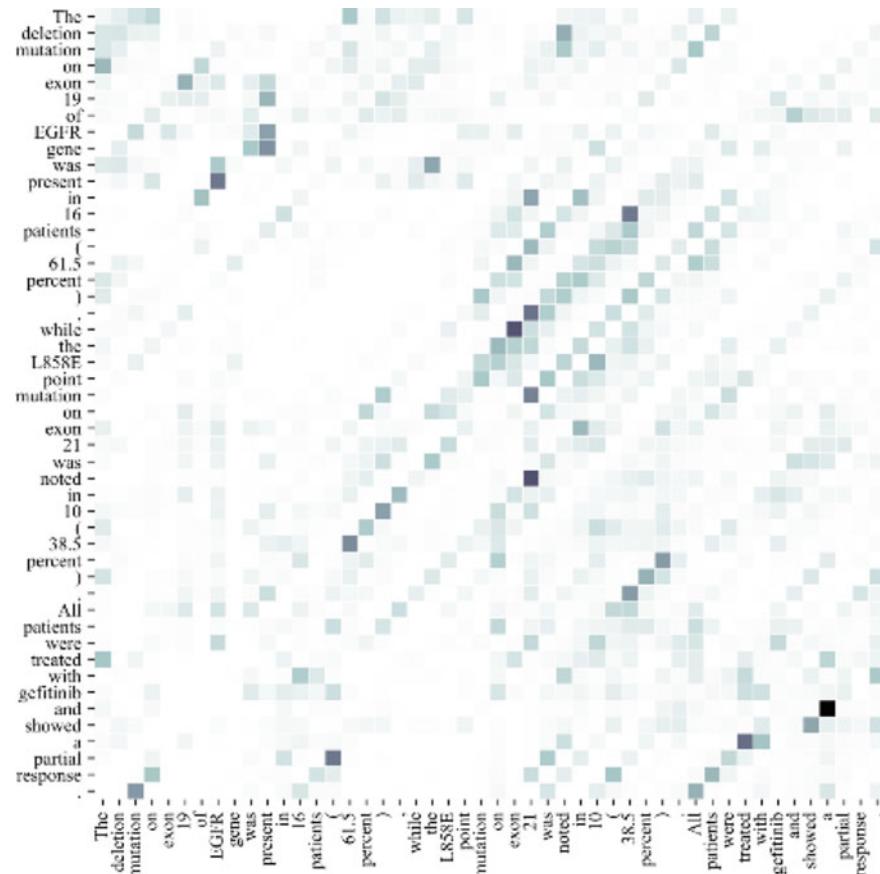
0.2 points. Similarly, layer normalization contributes 0.4 points. The entity-aware module and self-attention module contribute 0.5 and 0.7 points respectively, which illustrates that both layers promote our model to learn better relation representations. When we remove the feedforward layers and the entity representation, F1 score drops by 0.9 points, showing the necessity of adopting “multi-channel” strategy. We also notice that the BiLSTM layer is very effective for relation extraction, which drops the performance mostly (F1 relatively drops 2.3 points).

## 5.2 Interpretation of Self-attention

In order to intuitively interpret the strength of our proposed approach, we visualize the attention scores of self-attention layer to investigate whether the model has learned the crucial information that not exists in dependency tree of entities (i.e., *partial response*) for the relation extraction. In Fig. 4, we observed that our ESC-GCN focuses more attention on the center of the heat map, which means that higher scores are usually located in the middle of the sentence. Traditionally, words in the middle of the sentence are more likely to determine the relation of the entities. Besides, our model assigns the tokens (i.e., *showed a partial response*) related to entities relatively higher scores, which helps to predict the gold relation.

## 6 Conclusion

In this paper, we propose a novel neural model for relation extraction that is based on graph convolutional networks over dependency trees. By incorporating the context of the words related to entities with inter-dependencies of input sentence, our model can capture the long-distance dependency relation between target entities more effectively. Experimental results demonstrate that our model is superior to most baseline neural models. We further visualize the attention of our model to show how our relative position self-attention layer affects the model. In summary, our model effectively



**Fig. 4** Visualization of attention scores in the relative position self-attention layer

combines syntactic and semantic representations, which significantly improves the performance of relation extraction.

## References

1. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: International Conference on Computational Linguistics, pp. 2335–2344 (2014)
2. Zhang, D., Wang, D.: Relation classification via recurrent neural network. [arXiv:1508.01006](https://arxiv.org/abs/1508.01006) (2015)
3. Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., Jin, Z.: Classifying relations via long short term memory networks along shortest dependency paths. In: The Conference on Empirical Methods in Natural Language Processing, pp. 1785–1794 (2015)

4. Miwa, M., Bansal, M.: End-to-end relation extraction using lstms on sequences and tree structures. In: Annual Meeting of the Association for Computational Linguistics, pp. 1105–1116 (2016)
5. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: The Conference on Empirical Methods in Natural Language Processing, pp. 35–45 (2017)
6. Zhang, S., Zheng, D., Hu, X., Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Pacific Asia Conference on Language, Information and Computation, pp. 73–78 (2015)
7. Liu, Y., Wei, F., Li, S., Ji, H., Zhou, M., Wang, H.: A dependency-based neural network for relation classification. In: Annual Meeting of the Association for Computational Linguistics, pp. 285–290 (2015)
8. Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., Jin, Z.: Improved relation classification by deep recurrent neural networks with data augmentation. In: International Conference on Computational Linguistics (2016)
9. Peng, N., Poon, H., Quirk, C., Toutanova, K., Yih, W.t.: Cross-sentence n-ary relation extraction with graph lstms. In: Trans. Assoc. Comput. Linguist. **5**, 101–115 (2017)
10. Song, L., Zhang, Y., Wang, Z., Gildea, D.: N-ary relation extraction using graph state lstm. In: The Conference on Empirical Methods in Natural Language Processing, pp. 2226–2235 (2018)
11. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. In: The Conference on Empirical Methods in Natural Language Processing, pp. 2205–2215 (2018)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
13. Verga, P., Strubell, E., McCallum, A.: Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In: The Conference of the North American Chapter of the Association for Computational Linguistics, pp. 872–884 (2018)
14. Bilan, I., Roth, B.: Position-aware self-attention with relative positional encodings for slot filling. [arXiv:1807.03052](https://arxiv.org/abs/1807.03052) (2018)
15. Yu, B., Zhang, Z., Liu, T., Wang, B., Li, S., Li, Q.: Beyond word attention: using segment attention in neural relation extraction. In: International Joint Conference on Artificial Intelligence, pp. 5401–5407 (2019)
16. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: The Conference of the North American Chapter of the Association for Computational Linguistics, pp. 464–468 (2018)
17. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., S'eachdha, D.O., S.: Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: The International Workshop on Semantic Evaluation, pp. 33–38 (2019)
18. Nguyen, T.H., Grishman, R.: Relation extraction: perspective from convolutional neural networks. In: The Workshop on Vector Space Modeling for Natural Language Processing, pp. 39–48 (2015)
19. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Annual Meeting of the Association for Computational Linguistics, pp. 1556–1566 (2015)
20. Wu, F., Zhang, T., Souza Jr, A.H.d., Fifty, C., Yu, T., Weinberger, K.Q.: Simplifying graph convolutional networks. In: International Conference on Machine Learning (2019)
21. Santos, C.N.d., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. In: Annual Meeting of the Association for Computational Linguistics, pp. 626–634 (2015)
22. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Annual Meeting of the Association for Computational Linguistics, pp. 207–212 (2016)

# Overview on Job Running Times Prediction Algorithms for HPC Platform



Hao Wang and Yiqin Dai

**Abstract** The rapid development of high-performance computers has brought about a tremendous increase in computing power, but the data that needs to be processed is also multiplying. Improving the utilization of computing resources is a fundamental goal of high-performance computing systems. In high-performance computing systems, job scheduling systems often use backfill scheduling strategies to schedule jobs. This strategy is sensitive to the job running time. In the past, users' job running time was provided by users but was wildly inaccurate, which seriously affected computing resources. Therefore, it is necessary to improve the accuracy of job running time prediction. This article will introduce several types of existing job times prediction algorithms, but they all have some shortcomings. To improve the utilization of computing resources, the prediction algorithm should have high prediction accuracy, low underestimation, and more extensive scope of application.

**Keywords** High-performance computing · Backfilling scheduling · Job times prediction

## 1 Introduction

High-performance computing [1] has been widely used in the fields of science and engineering. The explosive growth of scientific computing demand provides a good growth environment for the rapid development of high-performance computing [2]. At the same time, it also proposes higher requirements for high-performance computing platforms. High-performance computing has the characteristics of strong computing power, fast processing speed, and strong scalability. Ensuring or even improving the preset service level agreement (SLA) [3] and making more reasonable use of high-performance platform computing resources are also significant issues.

On the high-performance computing platform, the job scheduling system is an essential part of it. It is responsible for scheduling the jobs submitted by users

---

H. Wang · Y. Dai

School of Computer Science, National University of Defense Technology, Changsha 410073, China

e-mail: [wanghao18d@nudt.edu.cn](mailto:wanghao18d@nudt.edu.cn)

according to a specific strategy. To improve the utilization of computing resources, the scheduling system will also adopt a backfilling strategy, that is, comprehensively considering the job's running time, scheduling the short-term jobs later in the job queue in advance, and filling the gaps in the running of each job. Job scheduling and backfilling are very dependent on the estimated running time of the job. In the past, the estimation of the running time of the job was provided by the user, but the work of Cirne et al. [4] showed that the estimated time of more than half of the job users is five times or more than the actual running time of the job. This behavior is very consistent with the psychology of the user. When the job is running, if the time allocated to the job has been exhausted and the job has not run to completion, the job will be killed. This situation is also called underestimation of the running time of the job. Many users may not know much about the job and the operating environment, but they don't want their submitted jobs to be killed. Therefore, they would instead provide jobs far beyond the required running time, but this approach is not beneficial for the system and will cause much waste of computing resources. Therefore, accurate prediction of job running time and consideration of underestimation issues help improve the utilization of computing resources of high-performance computing systems.

We divide the current algorithms for predicting job running times into two categories: analyzing source code to predict job running time and predicting job running time based on historical logs. Among them, the method of using historical logs is the mainstream research mode.

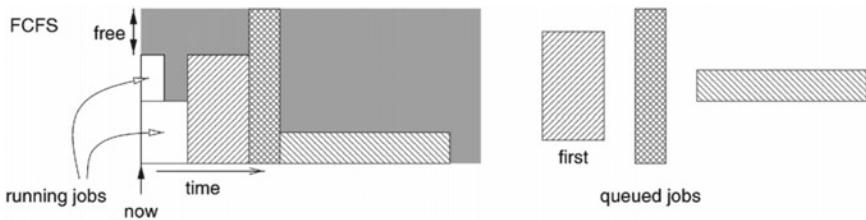
For this paper, the main contributions are as follows:

1. We explained from a microscopic perspective why job times with different prediction accuracy would have different impacts on the utilization of system resources.
2. We introduced several different types of job time prediction algorithms and analyzed their characteristics and shortcomings.
3. We look forward to the development trend of the prediction algorithm of the job times.

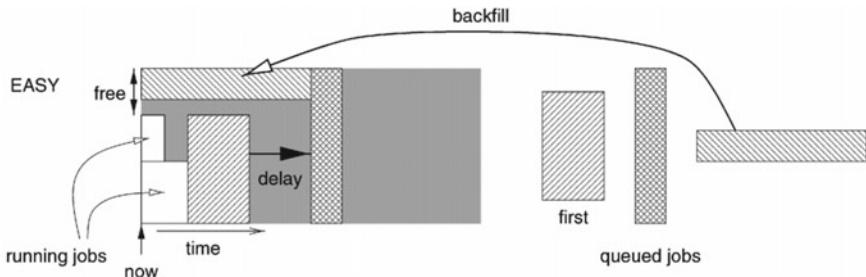
The article is organized as follows. Sect. 2 introduces the principle that backfill scheduling is sensitive to job running times, Sect. 3 introduces several algorithms for predicting jobs running times based on source code analysis and comparative analysis, Sect. 4 introduces several jobs running times prediction algorithms based on historical logs and analyzes their pros and cons, Sect. 5 concludes.

## 2 Backfill Scheduling Principle

In the job scheduling system, to further improve the utilization of computing resources, backfill scheduling is often used to schedule jobs. However, this method is sensitive to job running times, and assigning different sizes of running time to the same job will affect the scheduling sequence of each job by the operating system. Not



**Fig. 1** FCFS scheduling process



**Fig. 2** FCFS with EASY backfilling

only that, when the assigned job time is higher than its real running time, the extra time will be wasted, especially when the user provides this value. It will often be several times higher than the actual value. When the assigned job time is lower than its real running time, the consequences will be more serious, causing the job to fail and the job needs to be resubmitted. Take first-come-first-served (FCFS) and EASY backfill [5] as examples to introduce the scheduling process, as shown in Figs. 1 and 2.

In Fig. 1, there are two running jobs in the operating system. When the computing resource margin meets the demand of the first job in the queue, the three jobs in the job queue will be scheduled in the order of 1, 2, and 3 according to the FCFS strategy. In the FCFS algorithm, the job scheduling sequence is fixed, but the assignment time of the job affects when the job is scheduled. Inaccurate time allocation will lead to a waste of computing resources and a poor user experience.

In Fig. 2, the scheduling algorithm with EASY backfilling is shown. The EASY backfilling algorithm can schedule short-time jobs later in the queue in advance without affecting the scheduling of the first job in the job queue. Here, since the system can only meet the computing resource requirements of job 3 and the backfilling of job 3 will not affect the scheduling of job 1, job 3 is backfilled. According to the backfill strategy, the impact on job 2 will not be considered. It can be found that, compared to Fig. 1, the time when job 2 is scheduled is more affected by job time allocation, but the utilization of computing resources has been improved. At the same time, the reduction in average job waiting time also means an improvement in

user experience. Considering that the backfill schedule is sensitive to the job running time, it is necessary to predict the job running times accurately.

### 3 Backfill Scheduling Principle

Analyzing code structure to predict job running time is usually divided into two methods: static code analysis [6] and dynamic code analysis [7]. Freund et al. [6] first used the algorithm of static code analysis to predict the running time of the job. The central part of this algorithm is the code analyzer. The code is used as the analyzer's input, and the analyzer calculates the corresponding running time. In the analyzer, the code is divided into various code segments according to similar types. After dividing into segments, the instructions in each code segment are marked, which can be marked as the following types: vectorized instructions, parallel instructions, scalar instructions, and special instructions. Each type of instruction can also be subdivided, and the divided instructions will be respectively time-predicted and dispatched to the corresponding executor. This code division, marking, and prediction process will make time prediction more accurate and scheduling more effective. However, obviously, this instruction-dependent method is time-consuming and may not be acceptable to users. Due to conditional instructions, not all instructions are executed, so the method is less efficient.

Dirk et al. [7] considered conditional instructions and adopted branch prediction methods to predict the probability of instruction execution and then calculate the instruction's execution time. This is a dynamic code analysis method. The author uses a machine learning method to predict the running time of the code instructions. Compared with the static code analysis method, the prediction process is more intelligent, and the prediction is accurate. However, this method still takes a long time, and the user experience is not good.

In addition to static code analysis and dynamic code analysis, there are also job times prediction algorithms based on specific programming languages. Kiran et al. [8] proposed an algorithm for time prediction using compilation and code analysis based on R scripts. The algorithm extracts the source code of the R script and the critical parameters of the configuration file and simulates the execution of the R source code through the script file to predict the running time. The prediction work needs to use the token parser and the benchmark data set. The token parser marks the code, and the marked code is parsed line by line to determine the execution time of each line. The cumulative prediction time is the overall prediction running time. Susukita et al. [9] proposed an algorithm that can predict the running time of MPI programs in an MPI parallel environment and implemented it (BSIM) as an MPI program to run in the host system. When predicting the running time of the MPI program, BSIM intercepts the called MPI program to calculate the waiting and communication time of the program.

Wyatt et al. [10] proposed a more intelligent use of the convolutional neural network (CNN) algorithm PRIONN to predict the job's running time. PRIONN

also uses the job script as the input of the model. Due to the input requirements of CNN, the job script needs to be converted into a fixed-size image representation that CNN can recognize. Generally, the job script can be mapped into two structures, a one-dimensional sequence, and a two-dimensional array. When the job is mapped into a one-dimensional array, the information in the job script will be flattened into rows, and the character information will be mapped into pixels one by one. When the job is mapped into a two-dimensional matrix, the position structure of the original characters in the job script will be retained. The process of mapping job scripts into image representations must be performed on non-scheduler nodes to avoid interference with the scheduler. In PRIONN, when the job source code is mapped into a one-dimensional sequence, a one-dimensional CNN model is used to train the data. In this model, there are multiple one-dimensional convolutional layer-pooling layer groups and fully connected layers. When mapped to a two-dimensional matrix, a two-dimensional CNN model is used to train the data. In this model, there are multiple two-dimensional convolutional layer-pooling layer groups and fully connected layers. The output of the CNN model in PRIONN is a classifier. The final predicted job time is related to the nodes in the output layer. Each node in the output layer is mapped to a time value. The time values corresponding to all output layer nodes are accumulated to get the job running time. CNN can build a high-dimensional and very complex but powerful learning model. The job time prediction obtained from PRIONN will ultimately help the scheduler effectively schedule the job, thereby improving the utilization of computing resources.

In short, by analyzing the job source code, whether it is a fine-grained prediction of the time of each instruction or a coarse-grained job script as input, good prediction accuracy can be obtained. However, these methods have some common shortcomings. First, the underestimation of the prediction is not considered, and then the predicting process takes a long time. Most importantly, due to permissions, the source code of the job may not be available.

## 4 Predict the Running Time of the Job Based on Historical Logs

Using the job history log to predict the running time is the leading research model of current time prediction, mainly based on the similar job running time with similar jobs. The similarity here mainly refers to the similarity of the job feature attributes. The job attributes usually included in the history log is shown in Table 1. The estimate in Table 1 is the feature we hope to improve accuracy. It refers to the running time assigned to the job by the system specified by the user. When the actual running time of the job is less than this value, the extra time will be wasted, and once the actual running time of the job is more significant than this value, the job will be killed. So accuracy will affect the utilization of computing resources.

**Table 1** Common job attributes and their interpretation

Attributes	Meanings
CPU	The number of CPUs required for the job
Elapsed	The actual running time of the job
Estimate	User estimated job running time
Submit	The time when the user submitted the job
Wait	Job waiting time
User	User name
Job	Job name
T <sub>last2</sub>	The running time of the same job submitted in the last two times

#### 4.1 Use Statistical Methods to Predict Work Time

When predicting the job time, there was much work using statistical methods to study it. The more famous one is the Last-2 method proposed by Tsafir et al. [11]. This method is straightforward. It uses the attribute T<sub>last2</sub> shown in Table 1. For a job submitted to the queue, its predicted time is the average of the previous two submissions of the job. The accuracy of this method is minimal, because often the same job even in the case of the exact resource requirements such as CPU, memory, there will be differences in the value of other parameters closely related to the running time, resulting in a big difference in the actual running time.

Zhang et al. [12] used Hidden Markov Model (HMM) to propose a nonlinear and non-Gaussian time series prediction model. Before this, most of the time series prediction assumed a linear relationship between variables, such as the Box-Jenkins method [13]. Such assumptions would limit the application range of models. HMM is a method of predicting the running time of a job using the status of a log template. It assumes that the current job time depends on the time of the previous job. Each job time has multiple corresponding hidden states, and each hidden state represents a specific time slice. The final predicted time is the time required to transition from the initial state to the HMM absorption state.

In addition to the HMM, there is also a statistical model similar to it—Kalman filter [14]. The difference is that the HMM estimates the maximum posterior probability of time series predictions and uses noise sources in the prediction process, while the Kalman filter uses the least square error estimation and removes the noise source. Kalman filter is a recursive, iterative estimation algorithm. It first predicts the waveform, then observes the actual value, and finally combines the prediction and observation process to correct the waveform to obtain the final result.

When using statistical models to predict time series, it should be noted that they all assume that the job time is subject to a statistical distribution because they are based on the habit of user submission of jobs and require massive data support. The regularity of job submission habits in practice is not necessarily evident and compelling.

## 4.2 Predicting Running Time for Specific Types of Jobs

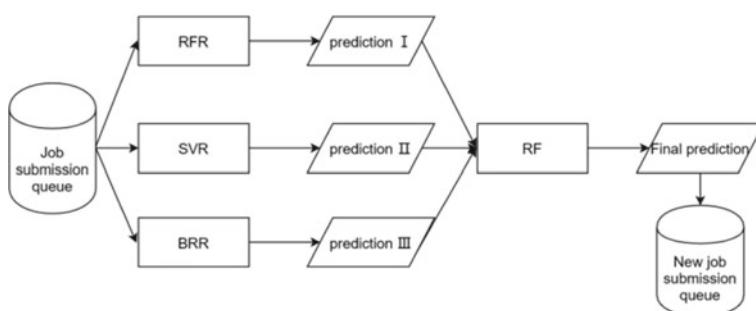
Predicting the running time of a particular type of job generally considers the large proportion of the job in the system. Research on the running time of this type of job will indeed help improve the utilization of the system's computing resources. In the study of WU et al. [15], the running time of the VASP job of the TC4600 cluster of the Supercomputing Center of the University of Science and Technology of China was analyzed in detail. This job accounted for 43% of the total number, and its machine time accounted for 46%.

When predicting the running time of a specific job VASP, the author extracted many unique attributes of the job. For example, KPOINTS and VOLUME represent the number of atoms inside the grain and the volume of the grain. It can be seen that these are all attributes that are closely related to the job itself, many of which are unique to VASP. For VASP, the acquisition of these attributes can help predict job time. Aiming at the prediction of VASP job time, the author proposes a model IRPA based on the secondary machine learning method, as shown in Fig. 3.

IRPA is a job time prediction model that combines regression and classification. This combination is also since more job attributes can be obtained for a specific job. Otherwise, it is challenging to implement. IRPA first gets three predictions through three sub-models, namely Random Forest Regression (RFR), Support Vector Regression (SVR), and Bayesian Ridge Regression (BRR), and the predicted values are used as the input of random forest classification (RFC). In RF, the target value is the closest input value to the actual value, which can be recorded as 0, 1, or 2. At the same time, we can get three probabilities from RF. The final result considers the predictions of the three sub-models. It can be described as the following formula:

$$PV = \frac{pv_1 * p_1 + pv_2 * p_2 + pv_3 * p_3}{p_1 + p_2 + p_3} \quad (1)$$

where  $pv$  is the predicted value of the sub-model, and  $p$  is the predicted probability of RF.



**Fig. 3** IRPA model

Such a secondary machine learning model comprehensively considers the three sub-models prediction results so that the prediction performance tends to be stable. However, as mentioned before, the IRPA model can only be suitable for specific job types. It requires a large number of job attributes to ensure the accuracy of the prediction results. If there are few attributes, the prediction results will be poor. At the same time, IRPA did not consider underestimation.

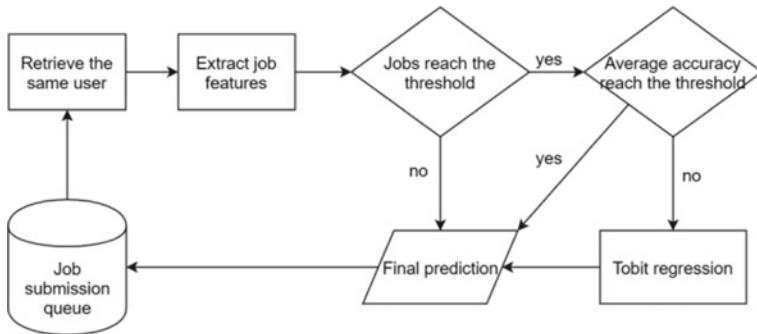
### **4.3 Meta-heuristic Optimization Algorithm Predicts Job Times**

The meta-heuristic optimization algorithm is based on the heuristic algorithm, which improves the heuristic algorithm. The algorithm is usually constructed by intuition or experience, and a feasible solution to the problem can be obtained under certain deviations. It includes an artificial bee colony algorithm (ABC), genetic algorithm, artificial neural grid algorithm, simulated annealing algorithm, etc. [16].

Pumma et al. [17] used the ABC algorithm and proposed an optimization algorithm with classification and linear regression technology based on a dynamic environment to predict job running time. The workflow is mainly divided into three steps: sampling, workload classification, and job time prediction. The workload is regarded as a kind of “black box” input. The algorithm does not care about the type of workload initially, but uses MICA and Perf for sampling, captures some system parameters of the load, and classifies the load according to the parameters, such as dense Linear algebra, N-body method, etc. Then carry out ABC algorithm model training for each type separately. This algorithm has achieved good prediction results, but if the job to be predicted does not belong to the already classified type, the model will fail.

### **4.4 Starting from the Data to Improve the Forecasting Effect**

There are usually three main parts when predicting the running time of a job: data, models, and prediction. Therefore, we can start the prediction work from these three aspects. Fan et al. [18] mainly started from the perspective of data processing, hoping to improve job time prediction accuracy by optimizing and filtering data. Much work before this has been devoted to improving the prediction accuracy by reducing the overestimation of running time but did not solve underestimating. As we know, underestimating time means disastrous consequences. Unfortunately, these two goals conflict, and improving the prediction accuracy of overestimation may increase underestimating. Fan Y et al. proposed an online adjustment framework TRIP. TRIP uses the data truncation capability of the Tobit model [19]. In TRIP, data that does not meet the conditions will be truncated so that more valuable data brings more accurate running time predictions. The TRIP model is shown in Fig. 4.



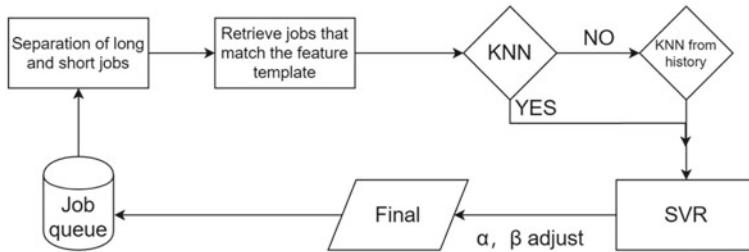
**Fig. 4** TRIP model

In TRIP, for each incoming user job, it first searches whether it is the same user, finds job sets with the same user name, the same project name, and the same job name counts the number of job sets and extracts data characteristics. For jobs that reach a certain number of job thresholds, the model will determine the historical prediction accuracy of its job set. Only when the number of jobs and the average prediction accuracy meet a specific requirement, the job to be predicted will use the prediction time submitted by the user. Otherwise, the model will submit the job to be predicted to the Tobit regression model. The Tobit model will filter out some data that does not meet the conditions according to the set conditions to reduce the underestimation rate while ensuring the prediction accuracy. It is a pity that TRIP does not perform well in improving accuracy and its prediction effects for different data sets are not the same.

#### 4.5 Predicting Job Times Based on Clustering Ideas

Predicting job times based on the clustering idea is more suitable for similar jobs with similar job running times. Based on the idea of clustering, Xiao et al. [20] proposed the GA-Sim prediction algorithm, which uses template similarity and KNN to obtain similar job sets, and then uses SVR to predict job running time. The predicting process is shown in Fig. 5.

In GA-Sim, the job queue is first separated from long and short jobs. This preliminary division reflects the idea of clustering, and the experiment also proves that this division is indeed effective. For the divided jobs, search for jobs that meet the feature template in the historical job collection. The author selects the user name, group name, queue name, and application name as the parameters of the feature template to filter the jobs that meet the conditions. Then, according to the numerical attributes of the job to be predicted, K neighbors of the job to be predicted are found from the set obtained in the previous step. In this way, the process of “clustering” is completed, and a set of similar jobs of the job to be predicted is obtained. When K



**Fig. 5** GA-Sim model

neighbors cannot be found, a certain number of neighbor sample points are obtained from the historical job. Finally, use the similar job set obtained to predict the job's running time with the SVR method. SVR is an excellent regression prediction algorithm with strong generalization performance and fast convergence speed. In order to avoid over-fitting the prediction results and to reduce the underestimation of the prediction, GA-sim introduces  $\alpha$  and  $\beta$  adjustment. Among them,  $\alpha$  adjustment uses the user's estimated time of job and uses the product of the two as the lower limit of the prediction, and  $\beta$  adjustment adds regularization. These two adjustment methods together make the prediction results more reliable. GA-Sim also takes into account the prediction accuracy and underestimation. The only shortcoming is that it is an online prediction method. The prediction model is trained in real-time, which may not be acceptable to users when it takes a long time.

## 5 Conclusion

This article summarizes and analyzes the relevant algorithms for predicting the job running time on the HPC platform. It can be divided into two categories: based on the source code to predict the job times and based on the historical logs to predict the job times. Analyzing the source code and predicting its running time can achieve better prediction results, but the most significant limitation of this type of method is the access rights to the source code, followed by the time-consuming prediction process and poor user experience. More researchers use historical logs to predict the running times, such as those using statistical methods, specific types of jobs, heuristics, optimized data, cluster-based, and so on. It makes good use of the advantages of its algorithms but has certain shortcomings. In general, an excellent job time prediction model should have the following three main characteristics: high prediction accuracy, low underestimation rate, and short prediction process. In addition, the generalization performance of the model and whether it is easy to deploy are also issues that should be considered.

## References

1. Ruibo, W., Kai, L., et al.: Brief introduction of Tian He exascale prototype system. *Tsinghua Sci. Technol.* **26**(03), pp. 113–121 (2021)
2. Liao, X.: MilkyWay-2: back to the world Top 1. *Front. Comput. Sci.* **8**(3) (2014)
3. Leff, A., Rayfield, J.T., Dias, D.M.: Service-level agreements and commercial grids. *IEEE Internet Comput.* **7**(4), 44–50 (2003)
4. Cirne, W., Berman, F.A.: comprehensive model of the supercomputer workload. In: *Proceedings of IEEE International Workshop on Workload Characterization*, pp.140–148 (2001)
5. Ahuva, M., et al.: Utilization, predictability, workloads, and user runtime estimates in scheduling the IBM SP2 with backfilling. *IEEE Trans. Parallel Distrib. Syst.* **12**(6), 529–543 (2001)
6. Freund, R.F.: Optimal selection theory for superconcurrency. In: *ACM/Ieee Conference on Supercomputing*. ACM (1989)
7. Tetzlaff, D., Glesner, S.: Intelligent prediction of execution times (2013)
8. Kiran, M., Hashim, A.H.A., Kuan, L.M., et al.: Execution time prediction of imperative paradigm tasks for grid scheduling optimization. *Int. J. Comput. Sci. Netw. Secur.*, pp. 155–163 (2009)
9. Susukita, R., Kimura, Y., Ando, H., et al.: Performance prediction of large-scale parallel system and application using macro-level simulation. *IEEE Press* (2008)
10. Graham, S., Park, C.: Assignment of dual port memory banks for a cup and a host channel adapter in an infiniband computing node [P]. US 6816889 B1 (2004)
11. Michael, R.W., Stephen, H., Todd, G., Adam, M., Dong, H.A., Michela, T.: PRIONN: Predicting Runtime and IO using Neural Networks, pp. 1–12 (2018)
12. Tsafrir, D., Etsion, Y., Feitelson, D.G.: Backfilling using systemgenerated predictions rather than user runtime estimates. *IEEE Trans. Parallel Distrib. Syst.* **18**(6), pp. 789–803 (2007)
13. Zhang, D., Ning, X., Liu, X.: Online prediction of time series based on RBF-HMM model. *J. Syst. Eng.*, 19–25 (in Chinese) (2010)
14. Box, G., Jenkins, G.M., Reinsel, G.C., et al.: Time series analysis: forecasting and control, 5th Edition. *J. Oper. Res. Soc.* **22**(2), 199–201 (2015)
15. Welch, G., Bishop, G.: An Introduction to the Kalman Filter (2006)
16. Gui-Bao, W., Yu, S., Wen-Shuai, Z., et al.: Runtime prediction of jobs for backfilling optimization. *J. Chin. Comput. Syst.* (2019)
17. Wang, Q.: Application of Metaheuristic Algorithm in Discrete Site Selection. Nanjing University of Aeronautics and Astronautics (in Chinese) (2010)
18. Pumma, S., Feng, W.C., Phunchongharn, P., et al.: A runtime estimation framework for ALICE. *Futur. Gener. Comput. Syst.* **72**, 65–77 (2017)
19. Fan, Y., Rich, P., Allcock, W. E., et al.: Trade-off between prediction accuracy and underestimation rate in job runtime estimates. In: *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE (2017)
20. Zhou, H., Li, X.: Tobit model estimation method and application. *Econ. Dyn.*, pp. 105–119 (2012) (in Chinese) (2012)
21. Xiao, Y., Xu, F., Xiong, M.: GA-Sim: a job running time prediction algorithm based on the combination of classification and case learning. *Comput. Eng. Sci.* (in Chinese)

# Effects of Game Perspectives Differences on Immersion Using Eye Tracking



Peng Li , Xu Jiang , and Xuebai Zhang 

**Abstract** The current development of science and technology has led to more and more expressions of video games. To solve the problem that whether the differences of game perspectives in video games will cause the difference of immersion, a method of measuring immersion by combining eye tracking and immersion scale was proposed. The setting of the experiment consisted of participants playing a game from one of the perspectives and recording their eye movement behavior, then filling out an immersion scale. The results showed that, in the absence of narrative guidance, players were more immersed in the third person than in the first person. In addition, the experimental data also showed that previous gaming experience had an impact on the difference in immersion caused by perspective. In future work, we will try more types of games and explore their differences in immersion.

**Keywords** Video games · Immersion · Eye tracking · First person · Third person

## 1 Introduction

With the advent of the technological age, the continuous development of video games has produced a new mode of human–computer interaction. Under this background, the visual performance of video games has been further developed, which has become a core topic discussed by game scientists and visual theory researchers. Research on computer vision has been a key subject of the world’s attention since the twentieth century, and these related research results are also an important foundation for related issues such as video games and visual expression. Among these researches, whether different game perspectives will affect player immersion is one of the key research topics at present.

Game perspective is one of the important elements in the design of video games. Currently, there are two mainstream game perspectives to meet the needs of players, namely the first person perspective and the third person perspective [1]. The first

---

P. Li · X. Jiang · X. Zhang ()

School of Computer & Information Engineering, Xiamen University of Technology, Xiamen 361024, Fujian, China

e-mail: [xbzhang@xmut.edu.cn](mailto:xbzhang@xmut.edu.cn)

person perspective means that the camera is located at the position of the controlled character's eyes, allowing the player to perceive the environment and the surrounding world through the character's eyes, while the third person perspective means that the camera dynamically following the controlled character according to the preset perspective and distance, allowing the player to observe the behavior of the controlled character. Daniel [2] discussed the visual engagement of players in third person perspective games, and the results showed that the third person perspective could create a stronger sense of action simultaneity than the first person perspective. Alena et al. [3] studied the impact of different preferences in the first person and third person on immersion in video games, and the questionnaire results showed that participants were more immersed in the first person than the third person in role-playing games and game preferences had no effect on the results. Thomas et al. [4] believed that the first person perspective is a natural way of human visual perception of the environment. Since the first person perspective is structured, more precise operations can be completed. However, a fixed perspective can make it difficult to see your surroundings or parts of your body. Video games overcome this disadvantage by having a third person perspective, which provides better vision and more control by being more aware of the environment in the game. Therefore, whether either of these two perspectives can make players more immersed in the virtual world of the game needs further research.

In game studies, scholars regarded immersion as an important part of game experience. At present, there are two main methods to measure immersion. One is qualitative measurement, such as interview method and questionnaire survey, etc., and the other is quantitative measurement, such as EEG and eye tracking. Ritu and Elena [5] divided immersion into five dimensions: time separation, focused immersion, high enjoyment, control and curiosity. Jennett et al. [6] compiled an immersive experience questionnaire based on the five dimensions of immersion based on previous relevant studies, which has become one of the more widely used immersion scales at present and is also used as a tool to measure immersion in playing video games. Diego et al. [7] used eye tracking to study the differences in people's sense of participation in VR games from different perspectives, and the experimental results showed that the first person perspective in VR games was more likely to cause discomfort, but it had a higher sense of immersion than the third person perspective, although the experimental data showed that the differences were small. In the past few decades, eye tracking has been applied to various fields. Gruden [8] analyzed the safety of people's behavior at traffic lights through eye movement behavior, and the experimental results showed that the use of digital devices has a significant impact on people's attention and a sharp decline in interest in other street elements. [http://apps.webofknowledge.com/OneClickSearch.do?product=WOS&search\\_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&colName=WOS&SID=6Em3iEsCcYsLN9YjxxU&field=AU&value=Gruden,%20CSome](http://apps.webofknowledge.com/OneClickSearch.do?product=WOS&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&colName=WOS&SID=6Em3iEsCcYsLN9YjxxU&field=AU&value=Gruden,%20CSome) people also used eye tracking to study how people inspect their vision and other related issues [9].

For this paper, the main contributions are as follows: try to use the combination of immersion scale and eye tracking data sampling to explore the difference between

immersion in the first person and the third person. The immersion scale can be used to understand the direct feelings of the subjects to the greatest extent, and the immersion level of the subjects can be obtained through the specific form of questionnaire. However, since the questionnaire relies on the subjective feelings and cognitive ability of the subjects, the use of the immersion scale alone may lead to deviation. Eye tracking can be analyzed by measuring physiological data to make up for the deficiencies of the immersion scale.

## 2 Theoretical Basis

### 2.1 Immersion

Immersion is a virtual feeling concept, and it is also a goal of entertainment art. For the definition of immersion, Csikszentmihalyi, an American psychologist, put forward the flow theory in the 1970s. He assumed that individual abilities and skills were taken as the X-axis and event challenges as the Y-axis to establish the coordinate system and set up the flow model. When the individual's abilities and skills are close to the challenges they are exposed to, they will enter a state of flow. Flow refers to a state where a person is fully engaged or involved in a certain activity. In this special state, the human ignores all irrelevant emotions and thoughts and enters a state of immersion [10].

Later, Csikszentmihalyi [11] improved the study of flow theory in the 1990s and put forward a three-channel model of flow theory, which is that the emergence of flow is related to the challenge difficulty of the activity and the skill level of the individual, and the best experience of the activity is when the challenge difficulty and the skill level of the individual reach a balance.

The flow theory was later applied to the field of games, and it was used by game designers to study the player's pleasure in the game, and strive to make the player obtain this pleasure through various methods. Jenova proposed Csikszentmihalyi's flow theory [12] that can be implemented in games. In order to maintain flow, he implements dynamic difficulty adjustment in the game, so that the difficulty and ability are always at a balance point.

### 2.2 Eye Tracking

Eye tracking refers to the tracking of eye movement by measuring the position of the eye fixation point or the movement of the eye relative to the head.

**Eye-Mind Hypothesis.** Research theories on eye tracking mainly come from the eye-mind hypothesis, which claims that changes in people's eyes can reflect

people's mental processes [13]. Under this hypothesis, researchers can observe the eye movement state of subjects to analyze the subjects cognitive process.

**Eye Tracker.** Eye tracker is a device that can measure the gaze position and movement information of the eyeball, which is widely used in the field of vision and psychology. As early as in the nineteenth century, eye tracker and its tracking technology were used to study people's mental activities by observing their eye movements, and to explore the correlation between eye movements and people's mental activities by analyzing the tracked eye movement behavior data [14]. Jennett et al. [5] believed that when players are in the immersive state, their attention will be more concentrated and the number of fixations will be less than that in the non-immersive state. When using eye tracking to study game immersion, they compared the number of fixations in immersive task and non-immersive task, and found that the difference was significant. The experimental results proved that eye tracking could be well applied to the study of game immersion.

### 2.3 *Game Perspective*

The choice of game perspective is a very important decision in game design. The current popular perspectives include the first person perspective, the third person perspective, the oblique 45 degree perspective, the overlooking perspective, etc. [1]. These different game perspectives can bring different game experiences.

**First Person Perspective.** The camera in a first person game is scheduled based on the eyes of the character you control, with a smaller field of view but a more detailed view of the environment. This kind of perspective is often used in shooting games, where most people don't see the character they're controlling.

**Third Person Perspective.** The third person perspective game is one of the most primitive game types. It is a game in which the player observes characters and actions in the scene from the perspective of an observer. This game perspective is usually above the game character controlled by the player. It conforms to people's visual habits, and is also the most suitable visual angle for people to observe the environment and game screens. So third person games were and are very popular with international game companies.

## 3 Experiment Design

### 3.1 *Experimental Instruments and Subjects*

Experimental Instruments: The environment of this experiment is the lowest equipment that can smoothly run the eye tracker. The main experimental environmental parameters are subject to this, so as to ensure the smooth operation of the eye tracking

device and the smooth operation of the experiment. The experimental environment is as follows:

1. Computer configuration: CPU: Intel (R) I7-8750H; RAM: 16 g; Graphics card: GTX1070; Operating system: Windows10 64bit; Monitor: 17.3 in.
2. Eye Tracker: Model: Tobii Eye Tracker 4C; Refresh rate: 90 Hz; Distance between eyes and screen: 60–80 cm.

Subjects: Thirty-four undergraduate and graduate students from Xiamen University of Technology were recruited randomly, including 2 invalid subjects who failed to collect data and 32 effective subjects, including 17 males and 15 females. They are between 19 and 30 years old. All subjects had normal visual acuity or corrected visual acuity, and had no color blindness or color weakness. Similar experiments had not been done before.

### ***3.2 Game Selection***

We chose games that could use both perspectives at the same time, especially games where the game mechanics didn't change significantly due to a change in perspective or presentation. We also looked for games that don't explicitly appeal to a particular gender. Based on this principle, we initially chose Minecraft and Grand Theft Auto V (GTAV), but later found that some of the participants were less receptive to the pixel style in Minecraft, so they ended up using GTAV. In this experiment, the first person and third person perspectives of "GTAV" will be used. In addition, the game has a lot of freedom and the experimental process does not involve the plot part.

### ***3.3 Experimental Preset Process***

Before the start of the experiment, we assumed that the immersion of the third person is higher than that of the first person without the guidance of the plot. And the Likert 7-level scale was used in this experiment.

Thirty-four undergraduate and graduate students from Xiamen University of Technology were randomly recruited. The preset process of each subject was as follows:

1. First, fill in the questionnaire to understand the game experience and game level of the subjects.
2. The subject sits at a distance of 60–80 cm from the eye tracker and the screen, with his eyes at the height of the center of the screen.

3. Record the subject number, perform head adjustment and eye tracking calibration, and inform the subject that the calibration is over when the accuracy is within the allowable error. In addition, the subjects were asked not to do too much movement as far as possible, so as not to affect the subsequent data collection.
4. Let the subjects draw the perspective number to decide the playing perspective. Before the experiment, the subjects were asked to enter the game in advance to familiarize themselves with the game operation, which lasted 15–20 min.
5. When the subjects think they can play without obstacles, they start the experiment formally and collect data. This process takes 15–20 min. In the whole experiment, there is no plot or text reading involved.
6. At the end of the experiment, the subject stops the game and closes his eyes. When data collection stopped, subjects were told to open their eyes. Fill out the Immersion Scale.
7. After the experiment, let the subjects receive the gifts.

## 4 Results Analysis

In general, both the eye tracking data and the immersion scale results support the original hypothesis that the third person perspective is more immersive than the first person perspective without the guidance of the plot.

The overall results of eye tracking data are shown in Table 1. The table shows the comparison between the number of fixations and fixation duration (unit: ms) under different perspectives.

Mean number of fixations per minute: T-test was used to get the result:  $t = 2.318 > t_{0.05}(25) = 1.708$ ,  $P = 0.01 < 0.05$ . Therefore, in the mean number of fixations, the mean number of fixations from the third person perspective was significantly less than that from the first person perspective.

Mean number of fixation duration per point: t test was used to get the result:  $t = 1.13 > t_{0.05}(30) = 1.697$ ,  $P = 0.13 > 0.05$ . Therefore, in the mean number of fixation duration per point, there was no significant difference between the third person perspective and the first person perspective.

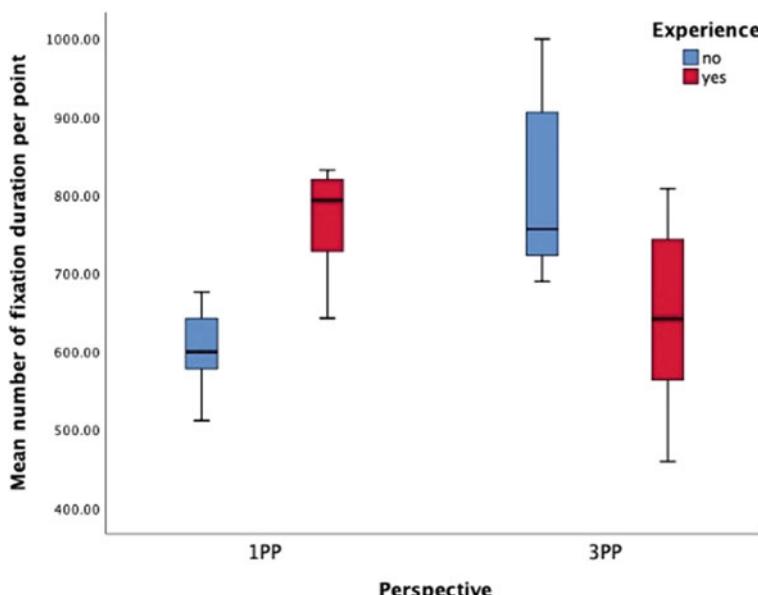
**Table 1** Comparison between the number of fixations and fixations on the market from different perspectives. 1

	Mean number of fixations per minute		Mean number of fixation duration per point	
	Mean	Std. Dev	Mean	Std. Dev
First person perspective	65.09	6.74	676.30	100.93
Third person perspective	57.58	11.07	726.11	144.17

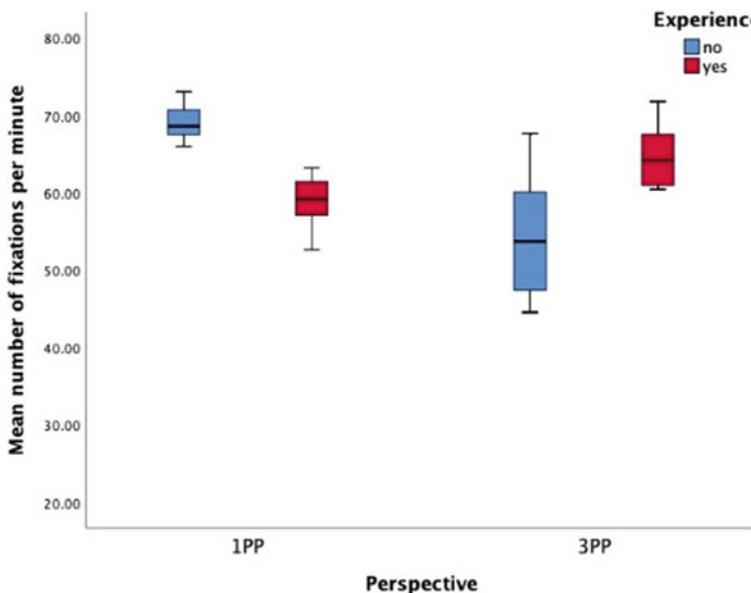
Then, two-way ANOVA is used to analyze the impact of previous game experience on immersion caused by perspective differences, and the following results can be obtained.

There was a significant difference in the mean number of fixations per minute between first person and third person perspectives ( $F(1, 28) = 5.526, P = 0.026$ ); However, the effect of player experience was not significant ( $F(1, 28) = 0.597, P = 0.446$ ); However, the interaction between these two factors was significant ( $F(1, 28) = 9.138, P = 0.005$ ). Among the 17 samples without gaming experience, the mean number of fixations per minute from the first person perspective (69.91) was significantly higher than that from the third person perspective (54.31), while among the 15 samples with gaming experience, the mean number of fixations per minute from the first person perspective (58.89) was slightly lower than that from the third person perspective (60.84). The specific data are shown in Fig. 1.

There was no significant difference in mean fixation duration per point between first person and third person perspectives ( $F(1, 28) = 1.357, P = 0.254$ ); Whether the player has experience or not has no significant effect on the mean fixation duration per point ( $F(1, 28) = 0.001, P = 0.973$ ); However, the interaction between these two factors is very significant ( $F(1, 28) = 22.001, P = 0.000$ ). Among the 17 samples with no gaming experience, the mean fixation duration per point from the first person perspective (606.6) was significantly lower than that from the third person perspective (806.93), while among the 15 samples with gaming experience, the mean



**Fig. 1** Mean number of fixations per minute



**Fig. 2** Mean number of fixation duration per point

fixation duration per point from the first person perspective (765.91) was higher than that from the third person perspective (645.29). The specific data are shown in Fig. 2.

At the end of the experiment, there were 32 complete immersion scales, including 16 for the first person perspective and 16 for the third person perspective. The immersion values under different perspectives are analyzed from six dimensions, and the results are shown in Table 2.

**Table 2** Comparison between the number of fixations and fixations on the market from different perspectives

	First person perspective		Third person perspective	
	Mean	Std. Dev	Mean	Std. Dev
Total immersion	99.38	20.72	113.56	18.1614
Basic attention	17.62	4.05	20.81	2.86
Temporal dissociation	15.75	3.92	18.38	4.62
Transportation	20.94	5.32	24.31	4.25
Challenge	12.81	3.27	14.19	2.51
Emotional involvement	12.31	3.07	13.81	3.33
Enjoyment	19.94	4.37	22.06	5.69

Total immersion: T-test was used, and the result was:  $t = 2.05 > t_{0.05}(29) = 1.699$ ,  $P = 0.02 < 0.05$ . Therefore, in total immersion, the total immersion score of the third person perspective was significantly higher than that of the first person perspective.

In summary, through relevant data analysis, it is found that the third person perspective and the first person perspective have significant differences in the mean number of fixations per minute in eye tracking data. Analyses of the experimental data suggest that in video games without plot guidance, the third person perspective is more immersive than the first person perspective. In addition, through two-way ANOVA of the immersion scale, experienced players were more immersed in the first person than inexperienced players, and inexperienced players were more immersed in the third person than experienced players.

## 5 Conclusions

Through a further understanding of immersion in games, it is helpful to design the virtual environment in games and design a more immersive visual representation, so as to improve the game experience of players and improve the efficiency of learning in games. In this article, we have raised the issue of immersion differences in playing video games (such as “GTAV”) from different perspectives. The differences were analyzed by the combination of immersion scale and eye tracking data. Based on the analysis of the collected samples, the following conclusions were drawn: in the absence of plot guidance, the subjects were more immersed in the third person perspective; in addition, whether they had previous game experience had an impact on the immersion of the difference in perspective. In the future, we will test more games, including more plot and interaction elements.

**Acknowledgements** This work was supported by the Xiamen University of Technology Scientific Research Starting Project for High-level talents (grant number YKJ19004R), and by the Xiamen University of Technology Postgraduate Science and Technology Innovation Program (grant number YKJCX2020101).

## References

1. Laramee, F.D.: Game Design Perspectives. Cengage Learning, Boston, MA (2002). [https://book.jd.com/publish/Cengage\\_Learning, Inc\\_1.html](https://book.jd.com/publish/Cengage_Learning, Inc_1.html)
2. Daniel, B.: Why can I see my avatar? Embodied visual engagement in the third-person video game. Journal **12**(2), 179–199 (2015)
3. Alena, D., Paul, C.: First person vs. third person perspective in digital games: do player preferences affect immersion? In: CHI’15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 145–148 (2015)

4. Thomas, K., Robin, B., et al.: Exploring the optimal point of view in third person out-of-body experiences. In: PETRA2016: Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp.1–4 (2016)
5. Ritu, A., Elena, K.: Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. *Journal* **24**(4), 665–694 (2000)
6. Charlene, J., Anna, C., et al.: Measuring and defining the experience of immersion in games. *Journal* **66**(9), 641–661 (2008)
7. Diego, M., Hai-Ning, L.: Evaluating merit, presence, and emulator sickness in VR games based on first-and third-person viewing perspectives. *J. Viewing Perspect.* E1830 (2018). <https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Liang%2C+Hai-Ning>
8. Gruden, C.: Safety analysis of young pedestrian behavior in signalized intersections: an eye-tracking study. *Journal* **13**(8), 4419 (2021)
9. Duchowski, T.: Eye Tracking Methodology: Theory and Practice. Springer, Secaucus, NJ (2003)
10. Mihaly, C.: Play and Intrinsic Rewards. Springer, Dordrecht (2014)
11. Mihaly, C.: Flow: The Psychology of Optimal Experience. Harper & Row, New York (1990)
12. Jenova, C.: Flow in game. *Journal* **50**(4), 31–34 (2007)
13. Just, M.A., Carpenter, P.A.: Eye fixations and cognitive processes. *Journal* **8**(4), 441–480 (1976)
14. Just, M.A., Carpenter, P.A.: A theory of reading: from eye fixations to comprehension. *Journal* **87**(4), 329–354 (1980)

# A Tolerance Classes Partition-Based Re-Definition of the Rough Approximations for Incomplete Information System



Lei Wang , Bin Liu , Xiangxiang Cai , and Chong Wang

**Abstract** The information with missing attribute value is known as the incomplete information system. In an incomplete information system, there exist tolerance relations satisfying reflexivity and symmetry among objects in the universe of discourse and the tolerance classes for each object can be derived from the tolerance relation directly. Generally speaking, the tolerance classes of all objects can form some coverages of the universe. The research of this paper is based on the tolerance classes of each object in the incomplete information system. Firstly, a method is proposed in which a partition of the universe is constructed by using tolerance classes of objects. Then, the upper and lower approximations of concept, i.e., a subset of universe, is re-defined on the basis of acquisition of the partition of the universe using proposed method and this new definition can improve the approximation accuracy of concept effectively compared to the traditional definition. Finally, an illustrated example is exhibited for demonstrating the method to obtain a partition to universe by tolerance class as well as the calculation of the rough approximations of concept based on the partition of universe by tolerance class.

**Keywords** Tolerance relation · Tolerance class · Partition · Lower approximation · Upper approximation

## 1 Introduction

The rough set theory, proposed by Pawlak [1, 2] nearly 40 years ago, is a well-known mathematics approach for data analysis, pattern recognition [3], knowledge update/discovery [4, 5], decision making [6, 7] and machine learning [8, 9]. The classical rough set theory is based on the partition by the equivalent class and it

---

L. Wang · B. Liu · X. Cai · C. Wang

Jiangxi Provincial Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing, Nanchang 330099, People's Republic of China  
e-mail: [2004992651@nit.edu.cn](mailto:2004992651@nit.edu.cn)

College of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, People's Republic of China

regarded that objects described by the same value on each attribute are indiscernible. All the indiscernible objects constitute an equivalent class or an elementary set. Any subset of universe of discourse, either being a union of some equivalent classes, is called crisp set, or being a rough set, which could be characterized by twain crisp sets, namely, lower approximation and upper approximations [1, 2, 10].

However, indiscernibility relation must have three properties, namely, reflexivity, symmetry and transitivity at the same time, this limits the real application of classical rough sets. The various extended rough set models have been proposed in succession through the various binary relation such as neighborhood relation, tolerance relation (reflexivity, symmetry) and preference relation (reflexivity, asymmetry and transitivity). A complete information system will turn into an incomplete information system owing to the missing of attribute value on some attributes. The definition of tolerance relation was proposed by Kryszkiewicz in incomplete information systems in 1998 [11, 12] and it was applied to relative attribute reduction. Leung and Li [13] proposed one computing approach for relative reduction of each object by employing maximal consistent block. Du and Hu dealt with approaches of reduction for attributes in an incomplete ordered information systems, in which certain attribute values maybe lost or absent [14].

For this paper, by utilizing tolerance classes of objects, we discuss re-definition of lower and upper approximations and their computation for incomplete information system. In Sect. 2, the fundamental concepts such as tolerance relation/class, maximal tolerance class as well as the definitions of low and upper approximations on the basis of similarity class are briefly introduced. In Sect. 3, the method for finding out one partition by the tolerance classes is discussed firstly. Then a novel re-definition of low and upper approximation of subset of universe is proposed on the partition by tolerance classes and their computational approach is given. By using this new definition of approximations, the approximation accuracy of concept can be improved compared to the existing definition of approximations which is based on the coverage of similarity classes. In Sect. 4, a numerical example is exhibited for illustrating the method to find out a partition to universe which is comprised of several tolerance classes and the calculation process of rough approximations of concept on the universe partition.

## 2 Preliminary

Some basic concepts, notations about incomplete decision information systems are outlined briefly [10–13].

A quadruple  $(U, C \cup \{d\}, V, f)$  denotes an information system,  $U$  and  $C$  are non-empty finite sets of objects and condition attributes, respectively.  $\{d\}$  is a decision attribute with  $C \cap \{d\} = \emptyset$ ;  $V = V_C \cup V_{\{d\}}$ , where  $V_C$  and  $V_{\{d\}}$  represent set of conditional attribute values and set of decision attribute values, respectively;  $f: U \times C \cup \{d\} \rightarrow V$  expresses mapping from  $U \times C \cup \{d\}$  to  $V$  such that each object from universe has one attribute-value in every attribute. Owing to lose or absent of attribute value on some

attribute, the decision information systems with missing value in some attributes is called incomplete decision information systems (for short, IDIS). Generally, the missing value is denoted as '\*' in IDIS, indicating that the values on certain attributes holds unknown.

In an IDIS,  $IDIS = \{ U, C \cup \{d\}, V, f \}$ ,  $AT \subseteq C$ , similarity relation  $SIM(AT)$  on  $U$  is defined as shown below [11, 12].

$$\begin{aligned} & SIM(AT) \\ &= \{(u, v) \in U \times U \mid \forall a \in AT, f(u, a) \\ &\quad = f(v, a) \text{ or } f(u, a) = * \text{ or } f(v, a) = *\} \end{aligned} \quad (1)$$

$R_{AT}(u) = \{v \in U \mid (u, v) \in SIM(AT)\}$  is called similarity class of object  $u$ . It is apparent that a certain object maybe belongs to at least two tolerance classes or to more. Therefore, the similar classes in an IIS are inherently overlapping.

$X \subseteq U$  is a subset of  $U$ , we call  $X$  a tolerance class w.r.t (with respect to)  $AT$  if  $\forall u, v \in X \rightarrow (u, v) \in SIM(AT)$ . If there is no subset  $Z \subseteq U$  such that  $X \subset Z$  and  $Z$  is a tolerance class w.r.t  $AT$ , then  $X$  is named as one maximal tolerance class of  $AT$  [10, 13]. The maximal tolerance class is the set of the maximum number of objects, in which the arbitrary two objects are similar. It is obvious that all the maximal tolerance classes can form a coverage of universe  $U$  [13].

The set of all the tolerance classes determined by  $AT$  is denoted as  $K_{AT}$  (except for singleton set) and the set of all the maximal tolerance classes by  $AT$  is denoted as  $T_{AT}$ .

Generally, the tolerance class which includes object  $x$  w.r.t to  $AT$  and the maximal tolerance class which contains object  $x$  with respect to  $AT$  are not unique.

To attribute set  $AT$ , the set of all tolerance classes which include object  $x$  and the set of all maximal tolerance classes which contain object  $u$  is denoted as  $K_{AT}(x)$  and  $T_{AT}(x)$ , respectively [13].

$$K_{AT}(x) = \{X \subseteq U \mid \forall u, v \in X, (u, v) \in SIM(AT)\} \quad (2)$$

$$T_{AT}(x) = \{X \in T_{AT} \mid x \in X\} \quad (3)$$

It is obvious that:  $K_{AT}(x) \subseteq K_{AT}$ ,  $T_{AT}(x) \subseteq T_{AT}$  and  $T_{AT}(x) \subseteq K_{AT}(x)$ .

**Definition 1** [11]  $IDIS = \{ U, C \cup \{d\}, V, f \}$  is an incomplete information system,  $AT \subseteq C$ . Lower and upper approximations on subset  $X$  of  $U$  are defined as followings:

$$\underline{Appr}_{AT}(X) = \{x \in X \mid R_{AT}(x) \subseteq X\} \quad (4)$$

$$\overline{Appr}_{AT}(X) = \{x \in X \mid R_{AT}(X) \cap X \neq \emptyset\} \quad (5)$$

### 3 The Tolerance Classes Partition-Based Definition of the Approximations

Generally, several tolerance classes can form a coverage of universe. Inspired by the partition of equivalent classes, we intend to partition the universe of discourse by usage of the tolerance classes owing to all the objects in one tolerance class are the same to attribute set  $A$ . The partition of universe by tolerance classes is investigated in this section and then a new definition of approximations of concept is proposed and the new definition can result in a bigger approximation accuracy of concept.

Since the tolerance relation need satisfy reflexivity and symmetry meanwhile, the tolerance relation can be characterized via an undirected graph  $G = (V, E)$ , among this  $G$ ,  $V$  being the set of vertices, which represent objects, and  $E$  being the set of edges connecting vertices among which there exist tolerance relation. Moreover, it is easy to verify that there exists at least one complete sub-graph of the undigraph  $G$ , which represents maximal tolerance class of all tolerance classes in universe  $U$ .

**Example 1** There is an incomplete information table, as shown in Table 1.  $U = \{u_1, u_2, u_3, u_4, u_5\}$ ,  $C = \{\text{att1}, \text{att2}, \text{att3}, d\}$ .

All the tolerance relations can be derived from Table 1 by Formula (1).

$SIM(AT)$

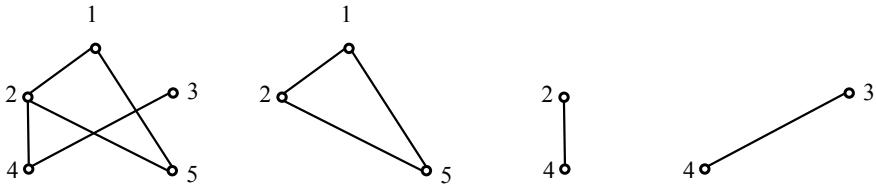
$$= \{(u_1, u_1), (u_2, u_2), (u_3, u_3), (u_4, u_4), (u_5, u_5), (u_1, u_2), (u_2, u_1), (u_1, u_5), (u_5, u_1), (u_2, u_5), (u_5, u_2), (u_2, u_4), (u_4, u_2), (u_3, u_4), (u_4, u_3)\}. \quad \text{Then we have:}$$

$$\begin{aligned} K_{AT} &= \{ \{u_1, u_2, u_5\}, \{u_2, u_4\}, \{u_3, u_4\}, \{u_1, u_2\}, \\ &\quad \{u_2, u_5\}, \{u_1, u_5\}, \{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\} \} \\ T_{AT} &= \{ \{u_1, u_2, u_5\}, \{u_2, u_4\}, \{u_3, u_4\} \} \end{aligned}$$

The graph of tolerance relation  $SIM(AT)$  and the corresponding complete sub-graphs of all its maximal tolerance classes, i.e.,  $T_{AT}$ , are as shown in Fig. 1.

**Table 1** An Incomplete information table

U	att1	att2	att3	d
$u_1$	*	0	*	1
$u_2$	0	0	4	1
$u_3$	3	0	1	*
$u_4$	3	0	*	1
$u_5$	1	*	4	1



**Fig. 1** A tolerance relation graph and the corresponding complete subgraphs of all its maximal tolerance classes

### 3.1 Partition of Universe by Tolerance Classes

Each tolerance relation in one tolerance class satisfies reflexivity, symmetry and transitivity simultaneously, so tolerance class is similar to equivalent class. Whether a number of tolerance classes can be found out from all the tolerance classes, so that these tolerance classes can constitute a partition of the universe  $U$ , namely:

$$\pi_{AT} = \{ X_k | X_k \in K_{AT} \wedge (\forall X_i, X_j \in \pi_{AT} \wedge i \neq j \rightarrow X_i \cap X_j = \emptyset) \wedge \bigcup_i X_i = U \} \quad (6)$$

It is obvious that:  $\pi_{AT} \subseteq K_{AT}$ .

For this reason, a method is examined, in which several tolerance classes are chosen from all the tolerance classes and they can constitute one partition of the universe. The process of the method is as following.

Firstly, the tolerance relation undirected graph  $G$  is obtained by the tolerance relation in an IDIS.

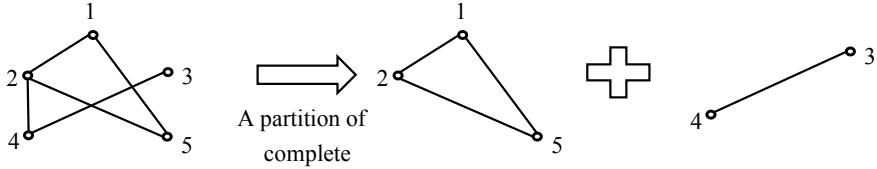
Secondly, the maximal complete subgraph  $g$  can be found out from the graph  $G$ : If there exist multiple maximal complete sub-graphs of the  $G$ , then select this one which contains the biggest number of vertices; If there exist multiple maximal complete subgraphs with the same number of vertices, then select one whose vertices associate the largest number of edges except for its own edges.

Thirdly, all the vertices and the edges which are associated with these vertices in the maximal complete subgraph  $g$  are removed from the graph  $G$  one by one. Thus, an updated tolerance relation undirected graph  $G'$  can be obtained.

Fourthly, a maximal complete subgraph can be found out over again from the graph  $G'$ , and so on, until the tolerance relation undirected graph becomes an empty graph.

**Example 2** (Continuation of Example 1). In IDIS of Example 3.1, one partition by tolerance classes to  $AT$ , denoted as  $\pi_{AT}$ , is not difficult to be found out in terms of the above-mentioned procedures, namely,  $\pi_{AT} = \{\{u_1, u_2, u_5\}, \{u_3, u_4\}\}$ .

The graph of tolerance relation  $SIM(AT)$ , one partition  $\pi_{AT}$  are as shown in Fig. 2.



**Fig. 2** A tolerance relation graph and its complete subgraphs corresponding to one partition

### 3.2 The Re-Definition of Approximations Based on Partition

A re-definition of the rough approximations of the subset of universe  $U$  is given on the basis of partition of universe  $U$  by tolerance classes.

**Definition 2**  $IIS = (U, C, V, f)$ ,  $AT \subseteq C$ ,  $X \subseteq U$ ,  $K_{AT}$  denotes set of all the tolerance classes determined by  $AT$  and  $\pi_{AT}$  (see Formula 6) represents one partition of the universe  $U$ .

$$\underline{appr}_{AT}(X) = \bigcup_{Z \in K_{AT} \wedge Z \subseteq X} Z = \{x | x \in Z \wedge Z \in K_{AT} \wedge Z \subseteq X\} \quad (7)$$

$$\overline{appr}_{AT}(X) = \bigcup_{Z \in \pi_{AT} \wedge Z \cap X \neq \emptyset} Z = \{x | x \in Z \wedge Z \in \pi_{AT} \wedge Z \cap X \neq \emptyset\} \quad (8)$$

**Property 1** In an  $IIS = (U, C, V, f)$ , for arbitrary  $X \subseteq U$  and  $AT \subseteq C$ , the following two formulae hold.

$$\underline{appr}_{AT}(X) \supseteq \overline{appr}_{AT}(X) \quad (9)$$

$$\overline{appr}_{AT}(X) \subseteq \overline{\overline{appr}}_{AT}(X) \quad (10)$$

**Proof** For  $R_{AT}(u) = \cup\{Z | Z \in K_{AT}(u)\}$ , we have  $\pi_{AT} \subseteq K_{AT}$ .

Suppose  $u \in \underline{Appr}_{AT}(X)$ , in terms of Formula (4),  $R_{AT}(u) \subseteq X$  holds. It follows that.

$R_{AT}(u) = \cup\{Z | Z \in K_{AT}(u)\} \subseteq X$ . So, for any  $Z \in K_{AT}(u)$ ,  $x \in Z$  and  $Z \subseteq X$ . Hence,  $u \in \underline{appr}_{AT}(X)$ . Therefore,  $\underline{Appr}_{AT}(X) \subseteq \underline{appr}_{AT}(X)$ .

We prove Formula (10):

$$\underline{Appr}_{AT}(X) = \cup\{R_{AT}(u) | u \in X\} = \cup\{\cup\{Z | Z \in K_{AT}(u)\} | u \in X\}$$

$$= \cup\{Z \in K_{AT} | Z \cap X \neq \emptyset\} \supseteq \cup\{Z \in \pi_{AT} | Z \cap X \neq \emptyset\} = \overline{appr}_{AT}(X)$$

Therefore,  $\overline{appr}_{AT}(X) \subseteq \overline{\overline{appr}}_{AT}(X)$ . This completes the proof.

**Example 3** (Continuation of Example 1). In IDIS of Example 1, let  $X = \{u_1, u_2, u_3, u_5\}$ . The approximations of subset  $X$  are computed according to Definition 1 and Definition 2 respectively.

$$\underline{Appr}_{AT}(X) = \{u \in X | R_{AT}(u) \subseteq X\} = \{u_1, u_5\}.$$

$$\overline{Appr}_{AT}(X) = \{u \in X | R_{AT}(X) \cap X \neq \emptyset\} = \{u_1, u_2, u_3, u_4, u_5\}.$$

$$appr_{AT}(X) = \bigcup_{Z \in K_{AT} \wedge Z \subseteq X} Z = \{u | u \in Z \wedge Z \in K_{AT} \wedge Z \subseteq X\} = \{u_1, u_2, u_5\}.$$

$$\begin{aligned}\overline{appr}_{AT}(X) &= \bigcup_{Z \in \pi_{AT} \wedge Z \cap X \neq \emptyset} Z = \{u | u \in Z \wedge Z \in \pi_{AT} \wedge Z \cap X \neq \emptyset\} \\ &= \{u_1, u_2, u_3, u_4, u_5\}\end{aligned}$$

So, the approximation accuracies of concept  $X$  in terms of Definition 1 is:

$$\mu_{AT}^{def1}(X) = \frac{|Appr_{AT}(X)|}{|Appr_{AT}(X)|} = 2/5 = 0.4.$$

And the approximation accuracy of concept  $X$  in terms of Definition 2 is:

$$\mu_{AT}^{def2}(X) = \frac{|appr_{AT}(X)|}{|appr_{AT}(X)|} = 3/5 = 0.6.$$

Among them,  $|\bullet|$  represents the cardinal of one set.  
Apparently,  $\mu_{AT}^{def1}(X) < \mu_{AT}^{def2}(X)$ .

The imprecision of a given concept is caused by the difference of its upper approximation and its low approximation, called the boundary region. The approximation accuracy for a concept  $X$  can indicate the size of its boundary region. The bigger the approximation accuracy of  $X$ , the little the boundary region and the imprecision of  $X$ .

## 4 The Algorithm for Acquisition of a Partition by Utilizing Tolerance Classes

The algorithm for acquiring one partition by using tolerance classes is constructed as follows. Moreover, an illustrate numerical instance is demonstrated to elaborate the application in computation of the low and upper approximations of subset of universe.

## 4.1 The Algorithm for Acquiring One Partition by Tolerance Classes

The algorithm for acquisition of one partition by using tolerance classes is as following.

**Algorithm 1.** The algorithm for acquisition of tolerance classes-based partition.

**Input:**  $IIS = (U, C, V, f), SIM(A), n = |U|$ .

**Output:** one partition  $\pi_A$  by tolerance classes.

**Step 1** To represent the tolerance relation undirected graph G by using  $n \times n$  order adjacent matrix.

**Step 2** If ( $n \neq 0$ ) Goto Step3; else Goto Step 10.

**Step 3** To find the maximal complete sub-graph g from the graph G.

**Step 4** If there exist multiple maximal complete sub-graphs of the graph G:  $g_1, g_2, \dots, g_m$  then select the one which contains the maximum number of vertices,  $g = \max_{\text{vertex}} (g_1, g_2, \dots, g_m)$ .

**Step 5** If there exist multiple maximal complete subgraphs with the same number of vertices:  $f_1, f_2, \dots, f_t$ , then select the one whose vertices associate the largest number of edges except for its own edges namely,  $g = \max_{\text{associated edges}} (f_1, f_2, \dots, f_t)$ .

**Step 6** To output the g and to record all the vertices  $m_i$  ( $i=1,2,\dots,n$ ) of the subgraph g.

**Step 7** To update graph G: for each vertex  $m_i$ , to delete all the edges associated with it and then to delete vertex  $m_i$ .

**Step 8** To update the value of  $n$ .

**Step 9** Goto Step 2.

**Step 10** The algorithm1 is finished.

## 4.2 Illustrative Example

Considering the following example, an IDIS is shown in Table 2,

$$\begin{aligned} U &= \{ i_1, i_2, \dots, i_{13}, i_{14} \}, AT = \{ at1, at2, at3, at4 \}, \\ X &= \{ i_1, i_2, i_3, i_4, i_7, i_8, i_9, i_{10}, i_{11}, i_{12}, i_{13}, i_{14} \}. \end{aligned}$$

1. The tolerance relations to attribute  $AT$  can be gotten from Table 2.

$$SIM(AT)$$

$$\begin{aligned} &= \{ (i_1, i_1), (i_2, i_2), (i_3, i_3), (i_4, i_4), (i_5, i_5), (i_6, i_6), (i_7, i_7), (i_8, i_8), (i_9, i_9), \\ &(i_{10}, i_{10}), (i_{11}, i_{11}), (i_{12}, i_{12}), (i_{13}, i_{13}), (i_{14}, i_{14}), (i_1, i_2), (i_1, i_3), (i_1, i_4), \\ &(i_2, i_3), (i_2, i_4), (i_3, i_4), (i_5, i_6), (i_5, i_7), (i_5, i_8), (i_5, i_9), (x_6, x_7), (x_6, x_8), (i_6, i_9), \end{aligned}$$

**Table 2** The Incomplete information table for illustrative example

U	at1	at2	at3	at4
$i_1$	3	2	0	*
$i_2$	*	*	0	1
$i_3$	3	2	*	1
$i_4$	*	2	0	1
$i_5$	*	5	*	3
$i_6$	2	5	0	*
$i_7$	2	*	*	3
$i_8$	*	5	0	*
$i_9$	2	5	*	3
$i_{10}$	1	*	*	0
$i_{11}$	*	2	*	0
$i_{12}$	3	2	1	*
$i_{13}$	*	2	1	1
$i_{14}$	3	*	*	1

$(i_7, i_8), (i_7, i_9), (i_8, i_9), (i_{10}, i_{11}), (i_{12}, i_{13}), (i_{12}, i_{14}), (i_{13}, i_{14}), (i_1, i_{11}), (i_2, i_6),$   
 $(i_2, i_8), (i_2, i_{14}), (i_3, i_{12}), (i_3, i_{13}), (i_3, i_{14}), (i_4, i_{14}), (i_8, i_{10}), (i_8, i_{14}), (i_{11}, i_{12}),$   
 $(i_2, i_1), (i_3, i_1), (i_4, i_1), (i_3, i_2), (i_4, i_2), (i_4, i_3), (i_6, i_5), (i_7, i_5), (i_8, i_5), (i_9, i_5),$   
 $(i_7, i_6), (i_8, i_6), (i_9, i_6), (i_8, i_7), (i_9, i_7), (i_9, i_8), (i_{11}, i_{10}), (i_{13}, i_{12}), (i_{14}, i_{12}),$   
 $(i_{14}, i_{13}), (i_{11}, i_1), (i_6, i_2), (i_8, i_2), (i_{14}, i_2), (i_{12}, i_3), (i_{13}, i_3), (i_{14}, i_3),$   
 $(i_{14}, i_4), (i_{10}, i_8), (i_{14}, i_8), (i_{12}, i_{11})\}$

2. According to the Algorithm 1, the following one partition by tolerance classes is obtained:

$$\pi_{AT} = \{ E_1, E_2, E_3, E_4 \}$$

$$E_1 = \{ i_5, i_6, i_7, i_8, i_9 \} .$$

$$E_2 = \{ i_1, i_2, i_3, i_4 \} .$$

$$E_3 = \{ i_{12}, i_{13}, i_{14} \} .$$

$$E_4 = \{ i_{10}, i_{11} \} .$$

3. The low and upper approximations of concept  $X$  in terms of Definition 2 is as follows.

$$\begin{aligned} \underline{app}_T(X) &= \bigcup_{Z \in K_{AT} \wedge Z \subseteq X} Z = \{i \mid i \in Z \wedge Z \in K_{AT} \wedge Z \subseteq X\} = E_2 \cup E_3 \cup E_4 \\ &= \{i_1, i_2, i_3, i_4, i_{10}, i_{11}, i_{12}, i_{13}, i_{14}\} \\ \overline{app}_T(X) &= \bigcup_{Z \in \pi_{AT} \wedge Z \cap X \neq \emptyset} Z = \{i \mid i \in Z \wedge Z \in \pi_{AT} \wedge Z \cap X \neq \emptyset\} = E_1 \cup E_2 \cup \\ &E_3 \cup E_4 = U. \end{aligned}$$

## 5 Conclusions

Incompleteness is a universal feature of information systems for various reasons in real applications. The principal motivation of this paper aims at constructing a partition of the universe by using tolerance classes in incomplete information systems and to redefine rough approximations of concept on the basis of the partition of the universe. Theoretical analysis and numerical examples show that this new definition has a finer performance in the approximation accuracy of concept than the previous definition. The dynamic update of the approximations according to the new definition while the object set varies is to be investigated in the further research.

**Acknowledgements** The work is supported by the fund of the Science & Technology Project of Education Department of Jiangxi Province (No. GJJ170995) and the National Science Foundation of China (No. 61562061).

## References

1. Zdzislaw, P.: Rough sets. *Int. J. Comput. Inform. Sci.* **11**, 341–356 (1982)
2. Zdzislaw, P., Andrzej, S.: Rudiments of rough sets. *Inf. Sci.* **177**(1), 3–27 (2007)
3. Yuhua, Q., Jiye, L., Pedrycz, W.: An efficient accelerator for attribute reduction from incomplete data in rough set framework. *Pattern Recognit.* **44**, 1658–1670 (2011)
4. Xuguang, C., Wojciech, Z.: Experiments with rough set approach to face recognition. *Int. J. Intell. Syst.* **26**(6), 499–517 (2011)
5. Jerzy, B., Salvatore, G., Roman, S.: Inductive discovery of laws using monotonic rules. *Eng. Appl. Artif. Intell.* **25**, 284–294 (2012)
6. Guangquan, Z., Zheng, Z., Jie, L., Qing, H.: An algorithm for solving rule sets-based bi-level decision problems. *Comput. Intell.* **27**, 235–259 (2011)
7. Huaxiong, L., Xianzhong, Z.: Risk decision making based on decision-theoretic rough set: a three-way view decision model. *Int. J. Comput. Intell. Syst.* **4**, 1–11 (2011)
8. Lin, S., Lanying, W., Weiping D., et al.: Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems. *Knowl.-Based Syst.* **192**(15), (2020)
9. Lin, S., Lanying, W., Jiucheng, X., et al.: A neighborhood rough sets-based attribute reduction method using Lebesgue and entropy measures. *Entropy* **21**(2) (2019)
10. Yan Yong, G., Hongkai, W.: Set-valued information systems. *Inf. Sci.* **176**(17), 2507–2525 (2006)
11. Marzena, K.: Rough set approach to incomplete information systems. *Inf. Sci.* **112**, 39–49 (1998)
12. Marzena, K.: Rules in incomplete information systems. *Inf. Sci.* **113**, 271–292 (1999)
13. Yee, L., Deyu, L.: Maximal consistent block technique for rule acquisition in incomplete information systems. *Inf. Sci.* **153**, 85–106 (2003)
14. Wensheng, D., Baoqing, H.: Dominance-based rough set approach to incomplete ordered information systems. *Inf. Sci.* **346–347**, 106–129 (2016)

# A Cyber Security Situational Awareness Extraction Method Oriented to Imbalanced Samples



Kun Yin , Yu Yang , and Chengpeng Yao

**Abstract** Due to the cyber security data contains a small proportion of attack data that cannot be effectively detected, and it is difficult for the traditional cyber security situation element acquisition model to extract accurate situation data from it. Therefore, this paper proposes a situational extraction method for uneven samples based on deep learning. First, use the CNN classifier to extract the characteristics of the cyber security data to obtain the classification accuracy of the original data set. Then, the deep convolution generation confrontation network (DCGAN) generates a uniform training data set on the basis of the original data set for small samples, and maps the network data to a two-dimensional matrix, which solves the problem of insufficient samples and sample imbalance; finally, Based on the balanced training data set, experiments are carried out on small samples through transfer learning. Experiments on the benchmark data set KDD'99 show that the data processing methods of transfer learning for small samples of R2L and U2R can obtain classification accuracy of 97.10% and 87.86%, respectively. Compared with the traditional model, the classification accuracy has been significantly improved.

**Keywords** Cyber security · Deep convolutional generative adversarial network · Transfer learning · Situation extraction

## 1 Introduction

### 1.1 A Subsection Sample

Situation extraction is a very important part of cyber security situation awareness [1]. Mature and complete situation extraction methods can prevent Internet-based attacks, misuse, or negligence. The situation extraction scheme of machine learning

---

K. Yin · C. Yao

Postgraduate Brigade, Engineering University of PAP, Xi'an 710016, China

Y. Yang ()

Academy of Information Engineering, Engineering University of PAP, Xi'an 710016, China

e-mail: [miaoyude@163.com](mailto:miaoyude@163.com)

has been developed to identify various legitimate network activities and potential threats [2]. The inability to correctly identify the feature distribution of different attack types results in insufficient generalization ability of the model and the sample cannot be correctly identified, and it is difficult to ensure the accurate acquisition of situation extraction.

At the same time, the design and implementation of a cyber security situation extraction scheme based on machine learning also face some challenges.

1. Traditional intrusion detection methods often use dimensionality reduction, compression and filtering techniques to eliminate detection noise when dealing with imbalanced multi-dimensional data points. Such a processing method tends to ignore the hidden important information when extracting the features of complex sample data, resulting in a higher false detection rate.
2. Deep learning usually requires a lot of labeled data to train general models. The pre-training data and test data of the neural network have different basic distributions. Therefore, it often takes a lot of work to collect and label training data. Especially under the current complex network and the replacement of a large amount of data generated by network operations, it is almost impossible to restart training the neural network to process data.
3. In order to ensure the efficiency and accuracy of situation extraction, it is necessary to preprocess the imbalanced data to improve the processing accuracy of small sample data. For example, repeated attacks are usually hidden in small network traffic [3]. If such attacks are not detected, the attacker can send a large number of offensive messages to the user, leaving hidden dangers in the attacked system. Therefore, the processing of small sample data is very important.

On the basis of theoretical experiments, this paper adopts the method of combining deep convolution generation adversarial network and migration learning to extract unlabeled imbalanced data from one or more source tasks in advance, and solves the problem of serious imbalance of training data. Then through transfer learning, based on the correlation between the source domain and the target domain, knowledge is transferred from the source domain to improve the detection performance of the model in the target domain.

In order to solve the problem of the lack of correct labeling of training samples, we use the deep convolutional generation confrontation network (DCGAN) [4] to generate satisfactory training data from the original samples, and at the same time, use the deep convolution generation confrontation network (DCGAN) to deal with small sample data Expand on the basis of the original data to get a balanced training set. Since DCGAN is more suitable for image processing, cyber security data needs to be pre-processed, and the Mahalanobis distance (Mahalanobis distance) is used to map the cyber security data into two-dimensional data. Therefore, DCGAN can be used to compress the measured threat samples.

The comprehensive detection accuracy of our proposed model on the KDD'99 dataset has reached 95.52%. In proportion, the detection accuracy of the other six popular learning methods [5] (SVM, random tree, random forest, NB tree, naive, Bayes and J48) on the KDD'99 data set is less than 94%. The rest of the structure

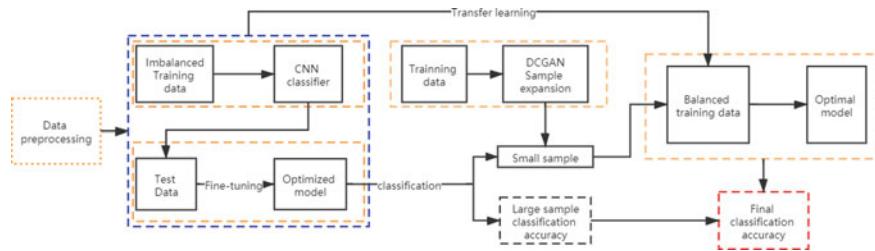
here is as follows. The second part introduces relevant background information and existing research solution. The fourth section discusses our experimental results. Finally, here is a summary in the fifth section.

## 2 Related Work

In recent years, the development of cyber security situational awareness has begun to be combined with machine learning. For example, Liu et al. [6] studied the element acquisition model based on Extreme Learning Machine (ELM), and Ding et al. [7] Using Support Vector Machine (SVM) as the feature acquisition model, Yang et al. [8] studied the feature acquisition model based on back propagation (BP, Back propagation). The above element acquisition model has the advantages of fast learning speed and good generalization ability; but when the amount of network data continues to increase, the accuracy and speed of element classification will be lost and reduced to varying degrees, and they all belong to a shallow structure and have the ability to represent limited. In order to solve the problem of insufficient fitting ability, deep learning has gradually replaced the dominant position of machine learning in the field of cyber security situational awareness. For example, some scholars have applied Convolutional Neural Network (CNN) to cyber security situation extraction [9] This method can better extract data features, it can effectively improve the classification accuracy and reduce the complexity of the model. Jiang et al. [10] input the paired standard (source + standard target) and generated (source + generated target) patterns into the discriminator, not only by distinguishing true and false to optimize, but also the input pair correctly classify the class label assignment. Seeliger et al. [11] create images similar to the presented stimulus image through a previously trained generator. Zhuang et al. [12] compared the transfer learning models and proved the feasibility of transfer learning and the importance of choosing the correct transfer learning model.

## 3 Method Implementation

The extraction of cyber security situation elements is essential to classify situation elements. The block diagram of the model is shown in Fig. 1. The improved CNN classifier model is used to learn sample features for all samples, and then according to the sample size and the classification accuracy obtained by the classifier Divide the sample into large samples (the amount of sample data is large, usually accounting for more than 10% of the total sample, the classifier index is superior), small samples (the number of samples and the number of large samples are quite different, which affects the accuracy of the classifier, usually accounting for The ratio is between 2 and 5% of the total sample, and the accuracy of the classifier is not ideal) and ultra-small samples (the amount of sample data is extremely small, usually accounting for about



**Fig. 1** The structure of the model proposed in this article

1% of the total sample, and the classifier cannot train meaningful classification, The classification accuracy is almost 0). Because the overall dimensions of the sample are imbalanced, relying on the strong fitting ability of the CNN classifier, a better classification accuracy of large samples has been obtained, but this approach will inevitably lead to the classification accuracy of small samples It is not ideal and needs further optimization. Through theoretical demonstration and experimental practice, specific measures are obtained: through the combination of DCGAN and transfer learning, first use the sample expander Deep Convolutional Generative Adversarial Networks (DCGAN) to train ultra-small samples The set is expanded to obtain a training set with balanced sample size, and then combined with migration learning to fine-tune the classifier to achieve the goal of achieving the same effect of small sample learning under the condition of short iterations and insufficient samples, and obtain a new adaptive small sample The CNN classifier improves the classification accuracy of small samples. This model is mainly divided into the following 4 steps:

Step ①: Data set preprocessing, processing the original one-dimensional network traffic data set into a two-dimensional matrix format suitable for the situational element acquisition and classification mechanism;

Step ②: Train the CNN classifier with supervised learning, and screen out small samples;

Step ③: Train this article to build a sample expander, enhance small sample data, and achieve sample balance;

Step ④: In the processed small sample training set, combine with transfer learning to fine-tune the classifier to improve the classification accuracy of the small sample.

### 3.1 Data Distribution

The classic KDD'99 is selected as the data set. In order to improve the accuracy of training and reduce the influence of useless data, only part of the KDD'99 data set is used as the experimental data. The data distribution is shown in Table 1. The data set includes attack data and normal data. The attack data includes DOS, probe, R2L and U2R, among which R2L and U2R are small samples. Select 20% of the attack data as the test data, and 80% as the training set.

**Table 1** Data quantity and proportion distribution

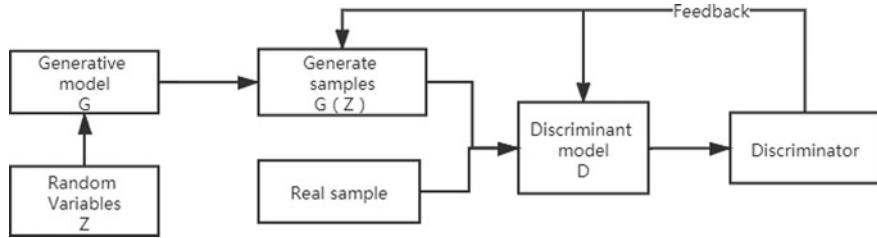
Sample type	Training set		Test set	
	Quantity	Proportion %	Quantity	Proportion %
Normal	60,593	57.945	14,188	56.172
Probe	30,422	29.092	7904	31.293
Dos	10,532	10.072	2348	9.296
R2L	2850	2.7254	762	3.016
U2R	172	0.16	56	0.221

### 3.2 Data Generation Model

The performance of deep neural networks depends on the quality and size of the labeled samples. In order to meet the training requirements of the deep learning model, small samples need to be expanded. The current mainstream generative models include DCGAN (Deep Convolutional Generative Adversarial Networks), WGAN (Wasserstein GAN), Least Squares GAN (LSGAN), etc. Among them, DCGAN has a relatively short training time and high accuracy, which is most suitable for network intrusion detection. Using DCGAN, we can obtain enough training samples to greatly improve the threat detection rate. The salient features of DCGAN can extract the correlation of high-dimensional data without labeling the target category. In addition, it can also automatically capture the distribution of sample data and generate training samples. It is usually difficult to capture positive samples (that is, samples in attack scenarios), because positive threat samples are usually submerged in a large number of normal data packets in reality. In this paper, DCGAN is used to solve the problem of sample imbalance, generate specific balanced samples, and expand small samples on the basis of real data. In the DCGAN model, the generator G generates fake samples close to the real samples to determine the discriminator, while the discriminator D filters out the fake samples generated by the generator G. In this process, the discriminator D is trained to improve the recognition ability, so that the generated sample is close to the real-world measurement value. Generation-antagonism The process of generating new samples realizes the dynamic balance of the system's high-dimensional non-convex continuous confrontation process (NASH equilibrium). By learning the existing attack samples, new attack samples can be generated, and as much information as possible is retained in the original sample data. The schematic diagram of sample generation and discriminant model is shown in Fig. 2.

Among them,  $G(z)$  represents the sample generated by the noise  $z$  of the generating network, and  $D(G(z))$  represents the probability of determining that the generated sample is a real sample after passing through the discriminating network. Finally, the generative model  $\min G(z)$  of the approximate optimal solution is obtained.

The output samples of the discriminant network are the probabilities from the real data, and are the parameters of the discriminant network and the generated network,



**Fig. 2** Schematic diagram of sample generation and discrimination model

respectively. The loss function of the generated network is shown in the following formulas:

$$L(G) = -E_{x \sim p_p(x)} \lg[D_\varphi(x)] \quad (1)$$

$$= -E_{z \sim p(z)} \lg[D_\varphi(G_\theta(z))] \quad (2)$$

$$= E_{z \sim p(z)} \lg[(1 - D_\varphi(G_\theta(z)))] \quad (3)$$

The core idea of DCGAN can be expressed in mathematical formula as shown in Eq. (4):

$$V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (4)$$

where  $x \sim p_{data}(x)$  means that  $x$  comes from the real distribution, and  $z \sim p_z(z)$  means that  $z$  comes from the simulated distribution.  $G$  represents a generative model, and  $D$  represents a classification model.

### 3.3 Mahalanobis Distance

In order to define the correlation between different features of the network flow feature vector, we use Mahalanobis Distance, as shown in Eq. 5:

$$\beta_{p,k}^j = \begin{cases} \sqrt{(W_p^j - W_k^j)^T S^{-1} (W_p^j - W_k^j)}, & t \neq p. \\ 0, & t = p \end{cases} \quad (5)$$

where  $w_{jp}$  is the value of the  $k$ -th feature of the  $j$ -th network flow feature vector,  $j = 1, 2, \dots, m$ ;  $k = 1, 2, \dots, m$ .

Meanwhile, in order to eliminate the interference caused by the correlation between variables, the covariance parameter is introduced to balance the probability between

the two categories, and the Mahalanobis distance is independent of the number of bits to convert the  $j$ th data stream into a symmetrical  $m$ -row  $m$ -column Hallow matrix  $E$ , as shown in Eq. 6:

$$E_{x_j} = \begin{bmatrix} \beta_{1,1}^j & \cdots & \beta_{1,m}^j \\ \vdots & \ddots & \vdots \\ \beta_{m,k}^j & \cdots & \beta_{m,m}^j \end{bmatrix} \quad (6)$$

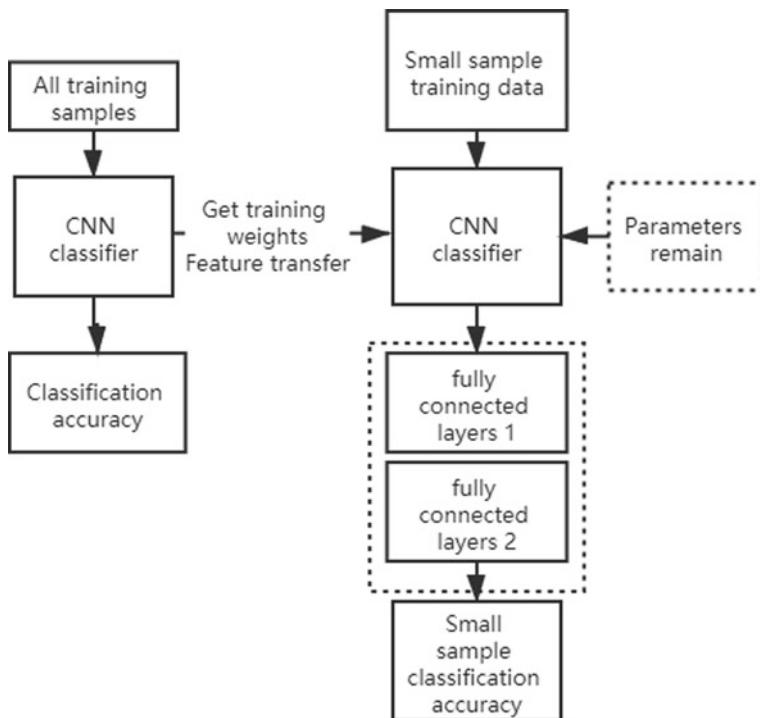
### 3.4 Small Sample Classification Method

For the small sample (smaller sample and expanded ultra-small sample) data set after the data set is expanded, transfer learning [13] can learn the internal characteristics of the sample well in the case of insufficient samples, and at the same time, the previously trained The CNN classifier has learned the internal features of large samples, and large and small samples have similar features, so a small sample classification method based on CNN classifier for migration learning is proposed. This method avoids the overfitting of small training samples and can effectively reduce the training time of the model. The migration learning process is shown in Fig. 3, which is divided into 3 steps:

1. Feature learning: Use CNN classifier to achieve sample feature learning.
2. Feature migration: use the weight parameters learned by the CNN classifier as the initial weight of the new classifier.
3. Classifier learning after migration: first freeze the first few layers of the CNN classifier network and remove the last layer, then add 2 layers of fully connected layers, fine-tune the last 2 layers through backpropagation, and get a new classifier suitable for small samples model.

In actual use, it is impossible to artificially identify large and small samples. A large sample that is misjudged by the CNN classifier as a small sample will be misjudged again in the new secondary classifier. Therefore, the classification accuracy of the small sample obtained by simulation is penalized for loss, and the specific formula is shown in Formula 7:

$$\begin{aligned} Acc_{smallsample}(True) = & Acc_{smallsample} - \frac{Size_{smallsample}}{\sum_1^2 Size_{smallsample}} \\ & \times [1 - (\sum_1^3 (Acc \times \frac{Size_{bigsample}}{Size_{allsample}}))] \end{aligned} \quad (7)$$



**Fig. 3** Transfer learning model

## 4 Experiment and Analysis

### 4.1 Model Evaluation Criteria

The selection of model evaluation criteria is very important. Accuracy, Precision, and Recall are all commonly used measurement indicators for element acquisition and classification problems. The samples are usually divided into positive samples and negative samples. Positive samples refer to samples that belong to the required class, and vice versa, samples that do not belong to the class are negative samples.

Among them, when TP is predicted to be a positive sample, the real is the same as the positive sample; FP is predicted to be a positive sample, and the real is a negative sample; FN is predicted to be a negative sample, and the real is a positive sample; TN is the predicted to be a negative sample, and the real is the same Negative sample.

$$\text{Accuracy} = \frac{TP + TN}{FN + TP + FP + TN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

Accuracy is usually selected as an indicator to measure the performance of a classification model, but when the sample distribution is imbalanced, it cannot accurately measure the overall classification accuracy of the model, and can only represent the classification accuracy of a large sample class. In engineering applications, Precision and Recall are usually negatively correlated. For further optimization, an optimization target needs to be selected. The selection criterion is that when the loss caused by missed judgment is large, Recall takes precedence. Conversely, when the loss caused by the misjudgement is greater, Precision takes precedence. Since this article solves the problem of acquiring cyber security situational elements, the omission of any attack in the security field may have serious consequences. In order to be able to accurately identify all attacks, I prefer to use recall rates in this article. The proposed prediction model is evaluated.

## 4.2 Data Set and Experimental Environment

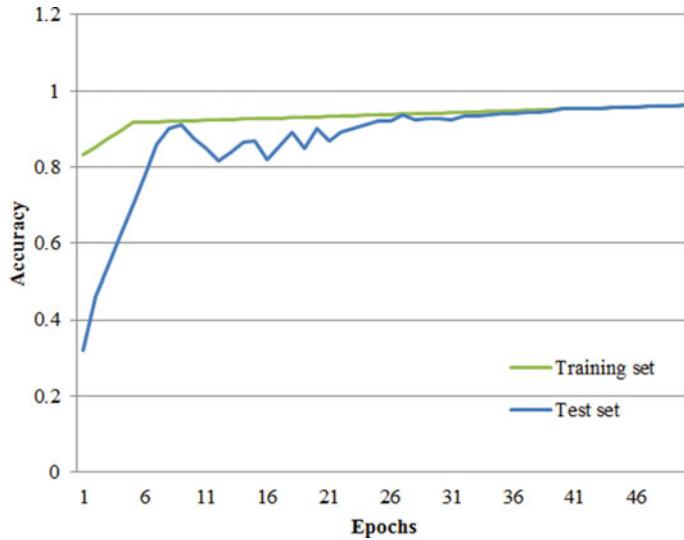
Experimental environment: Windows10 operating system, using Tensorflow to practice and test in python3.7 environment. The hardware configuration is: 64-bit operating system, the processor is Inter (R) CoreTM i7-7700HQ CPU 2.80GHZ, GPU is NVIDIA RTX2080TI.

In this experiment, KDD'99 [14] is selected as the test data set. The cyber security situation extraction model during the experiment is mostly a classification problem. The cross-entropy loss function is selected as the loss function, and the Adam [15] optimization algorithm is used to back-propagate the model. Set the number of epochs to 60 and batch\_size to 128. After setting the parameters, check the generalization ability of the model through the test set. The relationship curve between the accuracy rate and the number of iterations is shown in Fig. 4.

The accuracy rate increases as the number of iterations increases. When the number of iterations is about 45, the classification accuracy reaches 0.981, the loss function value gradually approaches 0, and the model becomes stable.

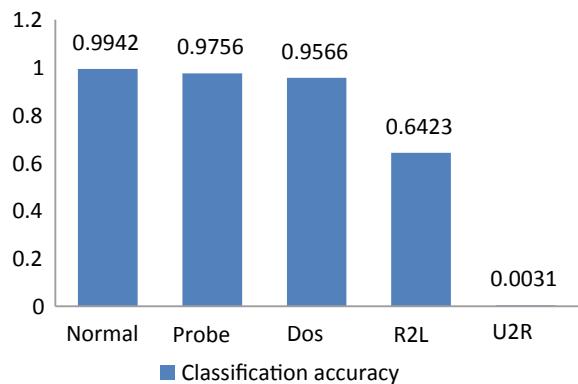
Meanwhile, in order to further obtain the specific classification accuracy of various factors of cyber security, other parameters are further calculated. As can be seen in Fig. 7, the classification accuracy of U2L and U2R is significantly lower than the other three categories, R2L and U2R. The sample size is small, and the features that can be extracted are also small, resulting in relatively low classification accuracy (Fig. 5).

The standard model has low classification accuracy for small samples. In order to improve the classification accuracy of the model for small samples, fine-tune the CNN classifier to obtain a new migration learning classifier. Select 80% of the R2L



**Fig. 4** The performance of classification performance and the number of iterations

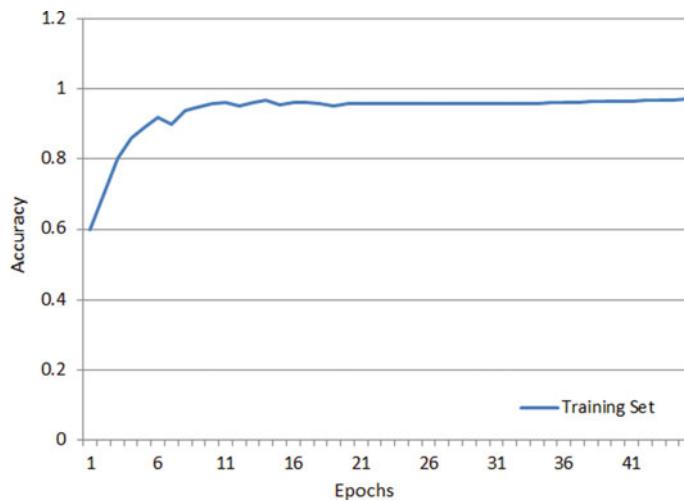
**Fig. 5** Classification accuracy of CNN classifier



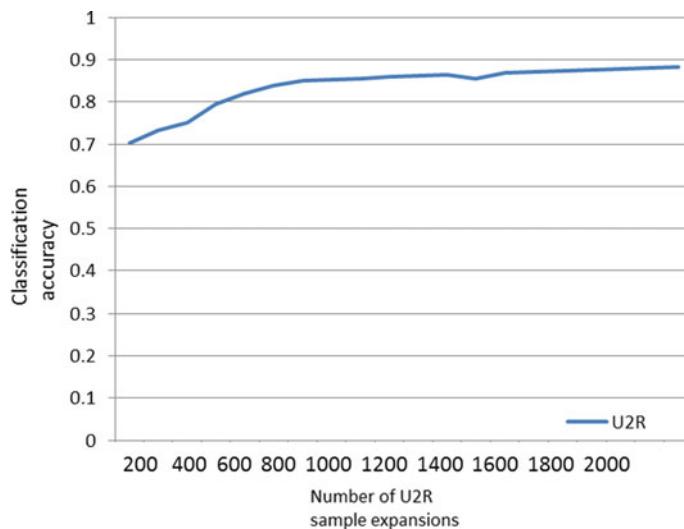
samples as the training set and 20% as the test set. The experimental results are shown in Fig. 6.

Analyzing experiments on R2L samples shows that as the number of iterations increases, the classification accuracy gradually increases. When the number of iterations reaches about 25, the model converges. Combined with the processing method of the transfer learning model, the classification accuracy increases by about 30%. In addition, the convergence speed of the model has been significantly improved, but the complexity of the model has not increased significantly.

The U2R samples are expanded by DCGAN, and then the expanded U2R samples are trained through the migration learning model, and the classification accuracy of the small samples is finally obtained as shown in Fig. 7.



**Fig. 6** The performance of transfer learning on the expanded sample training set



**Fig. 7** The performance of transfer learning on the expanded U2R sample test set

It is concluded through experiments that the size of the DCGAN sample expansion is positively correlated with the classification accuracy of the migration learning model. When the sample expansion reaches about 2000, the classification accuracy tends to be flat and close to the maximum. Compared with the unexpanded data set, the classification accuracy of U2R is improved by about 13%.

**Table 2** Table captions should be placed above the tables

Sample type	Attack type	SVM	CNN [16]	SRU-DCGAN [17]	PSO-DBN [18]	This model
Large samples	Normal	100	99.4	98.24	94.17	99.42
	Probe	91.0	83.5	96.72	83.45	97.56
	Dos	92.1	98.1	94.27	87.18	95.66
Small samples	R2L	80.4	20.61	93.26	84.40	97.10
	U2R	25.3	18.96	82.32	80.26	87.86

### 4.3 Model Comparison

Table 2 shows the comparison results between the proposed model and the classic algorithm and the data set KDD '99. The results show that the accuracy of this model is higher than that of SVM.

## 5 Conclusion

Experimental results show that our proposed model is better than traditional methods. On the one hand, a network data preprocessing method suitable for migration learning is established for complex and multi-dimensional network attack data. The model uses DCGAN to expand the samples and obtain enough Many training samples, and as much as possible to retain the information in the original sample data, provide high-quality data input for the migration learning model. On the other hand, the model combines migration learning based on the improved CNN classifier. Compared with the traditional model, the model obtains a faster rate and better initial performance, effectively learns the features of small samples, and improves the network. The classification accuracy of small samples in cyber security situational awareness improves the overall performance of the element acquisition model.

The purpose of transfer learning is to use the valuable information in one field to promote learning tasks in another field. Transfer learning between different models may be due to the lack of adaptability between latent spaces, and even when there are obvious differences between different latent spaces, it may lead to negative knowledge transfer. Therefore, it is very important to solve the computational complexity caused by feature migration. Next, we will do further research on the adaptability and robustness of the model.

**Acknowledgements** This work supported by Research on Key Technologies of Security Situation Awareness for Armed Police Force Optical Cable Network (WJY202130). Exploration and Practice of Online and Offline Mixed Teaching Mode of "Computer Network" Course (WJX2021120).

## References

1. Batsell, S., Rao, N., Shankar, M.: Distributed Intrusion Detection and Attack Containment for Organizational Cyber Security (2021)
2. Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial attacks and defenses in deep learning. *Engineering* **6**(3), 346–360 (2020)
3. Huang, N., Huang, S., Deng, Z.: Automatic detection of stack overflow attack in canary. In: 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), pp. 1418–1423. IEEE (2018)
4. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015)
5. Gupta, O., Raskar, R.: Distributed learning of deep neural network over multiple agents. *J. Netw. Comput. Appl.* **116**, 1–8 (2018)
6. Jinping, L., Jie-zhou, H., Tianyu, M., Wuxia, Z., Zhaojun, T., Pengfei, X.: Selective ensemble of KELM-based complex network intrusion detection. *ACTA Electronica Sinica* **47**(5), 1070 (2019)
7. Shufei, D., Bingjuan, Q., Hongyan, T.: An overview on theory and algorithm of support vector machines. *J. Univ. Electron. Sci. Technol. China* **40**(1), 2–10 (2011)
8. Yang, A., Zhuansun, Y., Liu, C., Li, J., Zhang, C.: Design of intrusion detection system for internet of things based on improved BP neural network. *IEEE Access* **7**, 106043–106052 (2019)
9. Zhang, Y., Chen, X., Guo, D., Song, M., Teng, Y., Wang, X.: PCCN: parallel cross convolutional neural network for abnormal network traffic flows detection in multi-class imbalanced network traffic flows. *IEEE Access* **7**, 119904–119916 (2019)
10. Jiang, H., Huang, K., Zhang, R., Hussain, A.: Style neutralization generative adversarial classifier. In: International Conference on Brain Inspired Cognitive Systems, pp. 3–13. Springer, Cham (2018)
11. Seeliger, K., Güçlü, U., Ambrogioni, L., GüclüTürk, Y., van Gerven, M.A.: Generative adversarial networks for reconstructing natural images from brain activity. *Neuroimage* **181**, 775–785 (2018)
12. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., He, Q.: A comprehensive survey on transfer learning. *Proc. IEEE* **109**(1), 43–76 (2020)
13. Zhang, X.S., Zhuang, Y., Yan, F.: Status and development of transfer learning based category-level object recognition and detection. *Acta Autom. Sin.* **45**(7), 1224–1243 (2019)
14. Kayacik, H.G., Zincir-Heywood, A.N., Heywood, M.I.: Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets. In: Proceedings of the Third Annual Conference on Privacy, Security and Trust, vol. 94, pp. 1723–1722 (2005)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
16. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(9), 2805–2824 (2019)
17. Yang, J., Li, T., Liang, G., He, W., Zhao, Y.: A simple recurrent unit model based intrusion detection system with dcgan. *IEEE Access* **7**, 83286–83296 (2019)
18. Wei, P., Li, Y., Zhang, Z., Hu, T., Li, Z., Liu, D.: An optimization method for intrusion detection classification model based on deep belief network. *IEEE Access* **7**, 87593–87605 (2019)

# Mine Cable Fault Distance Detection



Zezhong Liu Ming Lu Zuguo Chen Wang Cheng and Jinyu Wang

**Abstract** In a modern mine, the transmission line is an important part. When the transmission line fails, it will cause damage to the entire transmission system. How to quickly and accurately find the fault point and eliminate the fault is an important issue. This paper analyzes the fault characteristics of different fault types, and uses traveling wave theory to carry out fault location. Phase-to-modulus transformation is used to eliminate the inter-phase coupling between three phases. After phase selection, for the traveling wave of the faulty phase, wavelet transform is used as a mathematical analysis tool to perform wavelet transformation on the fault traveling wave. The high-frequency part decomposed by wavelet transform is modulated. Then, the high frequency part of traveling wave decomposed by wavelet transform is processed by modulus maximum. The fault traveling wave head is extracted, and the time points of the first arrival and the first reflection wave head are obtained. Combined with the traveling wave propagation theory, the location of the fault point is calculated. In this paper, using the power system module library in MATLAB software, a system model of a transmission line is established for fault simulation for different fault types and fault distances, and then the fault data is processed and analyzed to verify the accuracy of the phase selection method and the distance measurement method.

**Keywords** Fault location · Wavelet transform · Traveling wave theory

## 1 Introduction

With the widespread use of cables in underground mines, the difficulties in power system fault diagnosis have become more severe. Due to the humid environment of the mine, the cables have been corroded for many years, and many cables have begun to enter old age. Part of the cable line has already experienced insulation aging faults due to the early investment time. With reference to the general law of failure development, the probability of cable failure should conform to the Lobine curve [1],

---

Z. Liu · M. Lu ( ) · Z. Chen · W. Cheng · J. Wang

School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China

e-mail: [mlu@hnust.edu.cn](mailto:mlu@hnust.edu.cn)

that is, the failure rate in the early and late stages of the entire service life is higher, and the failure rate in the middle period is lower. Due to the influence of uncertain factors such as underground gas in the mine, emergencies caused by cable failure will bring serious threats to life and property safety, and even cause severe social impact.

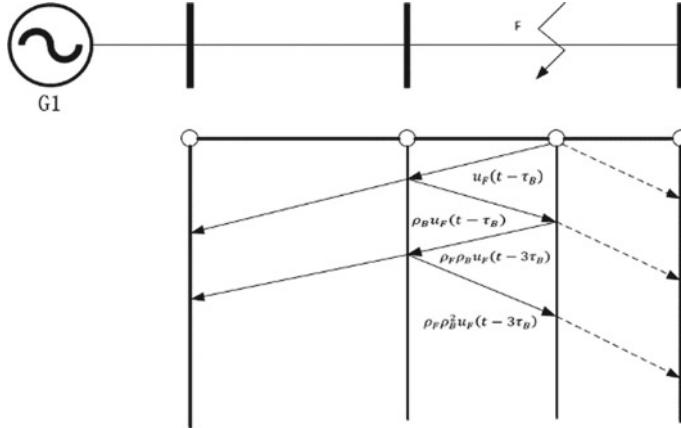
At present, line fault location includes impedance method and cable fault location method based on high-speed photoelectric sensing technology. Among them, impedance method includes classical bridge method and distributed parameter calculation high-resistance fault method. This type of method is used when the phase difference is small. There will be a large error in the bottom. The cable fault location method of high-speed photoelectric sensing technology uses optical cable communication between the photo-magnetic sensor and the photoelectric converter. For long lines, the investment in the optical cable is too large and it is not suitable for engineering [2]. The wavelet transform used in this article has the ability to characterize the characteristics of signal mutations and has good processing effects on non-stationary signals [3]. It can analyze and suppress the interference of the results of wavelet transform of signal at different scales, extract the signal fault characteristic parameters, and realize the accurate fault location.

## 2 Fault Detection Method Based on Traveling Wave

### 2.1 Traveling Wave Ranging

Assuming that the line is a lossless line, that is, the resistance and conductance are both 0, and a ground fault occurs at point F of the transmission line at time  $t = 0$ , and the incident voltage at the fault point is  $u_F(t)$ , When the radio wave reaches the fault point, reflection and refraction occur. At the time  $t = \tau_B$ , the fault traveling wave reaches the end of the bus bar, reflection and refraction occur, and the time  $t = \tau_B$  is delayed. Denoted as  $u_F(t - \tau_B)$ , set the reflection coefficient at the bus bar B end to  $\rho_B$ , then the reflected voltage wave reflected by the fault traveling wave arriving at the bus bar B is  $\rho_B u_F(t - \tau_B)$ , and the reflected voltage wave is the same wave speed propagates towards the fault point F, after a delay time  $\tau_B$ . After reaching the fault point F, reflection and refraction occur again at the fault point F, assuming that the reflection coefficient is  $\rho_F$ , Then the secondary reflection voltage at the fault point F is  $\rho_F \rho_B u_F(t - 3\tau_B)$  [4, 5] (Fig. 1).

When a single-phase grounding fault occurs on a transmission line, the faulty phase generates a traveling wave, while the non-fault phase also generates a traveling wave due to the influence of mutual inductance. For this reason, it is necessary to extract the traveling wave information from the three-phase voltage and current waveforms. The three-phase voltage and current are multiplied by the Karen Bell transformation matrix or matrix to transform them into linear modulus components  $\alpha$  component,  $\beta$  component and zero modulus component 0 component. The three



**Fig. 1** Traveling wave transmission network diagram described by grid method [6]

modulus components decomposed after the traveling wave undergoes phase-modulus transformation. Since the wave speed of the line modulus component is greater than the wave speed of the zero modulus component, the zero modulus component reaches the bus end after the line modulus component, and only when a ground fault occurs in the line A zero-modulus component will be generated, but no modulus component will be generated when a non-ground fault occurs [7]. This feature can be used as an important criterion for determining whether it is a ground fault in the fault phase determination. This article uses Karen Bell transformation moments:

$$\begin{bmatrix} u_\alpha \\ u_\beta \\ u_0 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} u_a \\ u_b \\ u_c \end{bmatrix} \quad (1)$$

A short-circuit fault occurs on the transmission line at point F, and the location of the fault point is determined by detecting the time difference between the two reflected waves reaching the bus:

$$x = \frac{v \times (t_2 - t_1)}{2} \quad (2)$$

where  $t_1$  is the time when the reflected wave first arrives at the bus bar,  $t_2$  is the time when the reflected wave arrives at the bus bar for the second time, and the wave velocity is determined by the characteristics of the medium [7].

$$v = \sqrt{\frac{1}{LC}} \quad (3)$$

## 2.2 Wavelet Transform

Suppose the function  $f(t) \in R$ . The singularity of the function  $f(t) \in R$  at a certain place is generally characterized by the singularity index lipischitz  $\alpha$ . The definition of the index is: Let  $0 \leq \alpha \leq 1$ , if there is a constant  $K$  at the point  $t_0$ , for the neighborhood  $t$  of the point  $t_0$ , the following formula holds (4),

$$|f(t) - f(t_0)| \leq K|t - t_0|^\alpha \quad (4)$$

Then we say that  $f(t)$  is lipischitz  $\alpha$  at point  $t$ . If  $\alpha = 1$ , it is said that  $f(t)$  is differentiable at  $t_0$ , that is,  $f(t)$  has no singularity; if  $\alpha = 0$ , it means that  $f(t)$  is discontinuous at  $t_0$ . The larger the  $\alpha$  is, the closer the  $f(t)$  is to the rule. On the contrary, the smaller the  $\alpha$  is, the sharper the change of  $f(t)$  at the  $t_0$  point [8].

The singularity of a function can be characterized by its lipischitz  $\alpha$ , and the magnitude of the value can be obtained by calculating the maximum value of the wavelet transform. The definition of the modulus maximum is given below: When the scale is  $2^j$ , for  $t$  in the neighborhood of  $t_0$  [9].

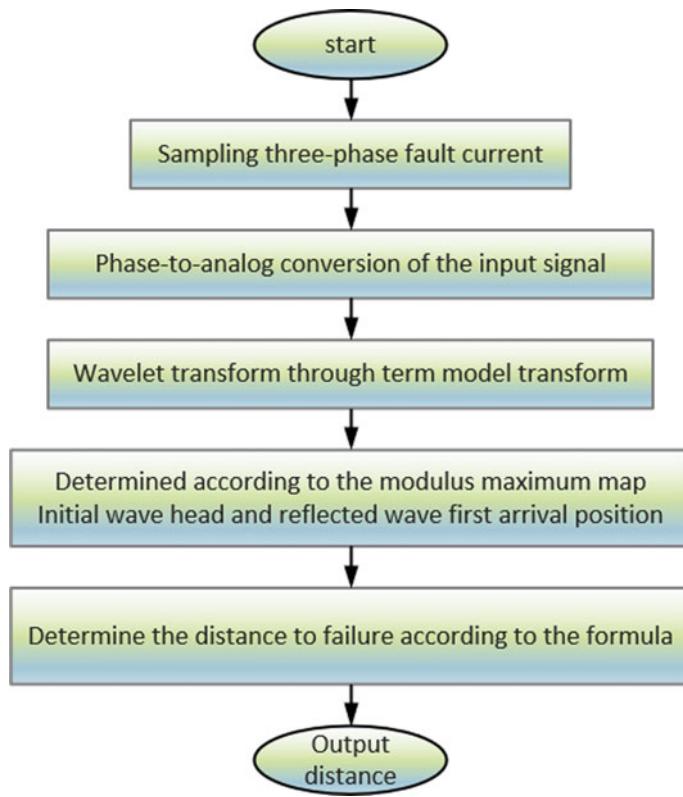
$$|W_{2^j}f(t)| \leq |W_{2^j}f(t_0)| \quad (5)$$

In the wavelet transform, the point  $t_0$  that satisfies the Formula (5) is called the modulus maximum point, and  $W_{2^j}f(t_0)$  is called the modulus maximum [10]. The singularity detection theory is aimed at the sudden change of the signal, and its sudden change time and the degree of change are described by mathematical theory. When using wavelet transform to analyze the signal, because the signal can be localized in the time-frequency domain, and the width of the time window and frequency window can be automatically adjusted, so that it can detect that the abrupt signal increases with the change of the scale. The maximum value point of the modulus caused by noise will be reduced, and the maximum value point of the transformed coefficient modulus caused by the fault is more obvious. Therefore, we can find out the maximum value point of the modulus [11, 12].

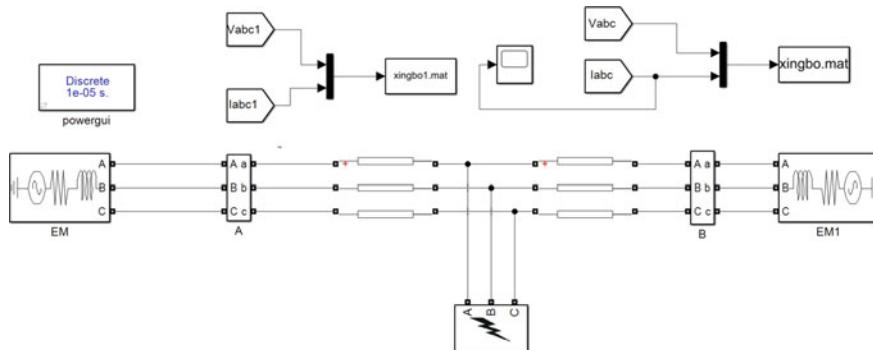
## 3 Experimental Verification

This article takes the 200 km, 110 kV high-voltage line as an example, and uses MATLAB's power system toolbox for modeling and simulation. The total length of the line is 200 km, of which the total length of the first section of the line is 100 km, and the total length of the second section of the line is 100 km. The flow chart of traveling wave fault location distance protection based on wavelet analysis is shown in the Fig. 2.

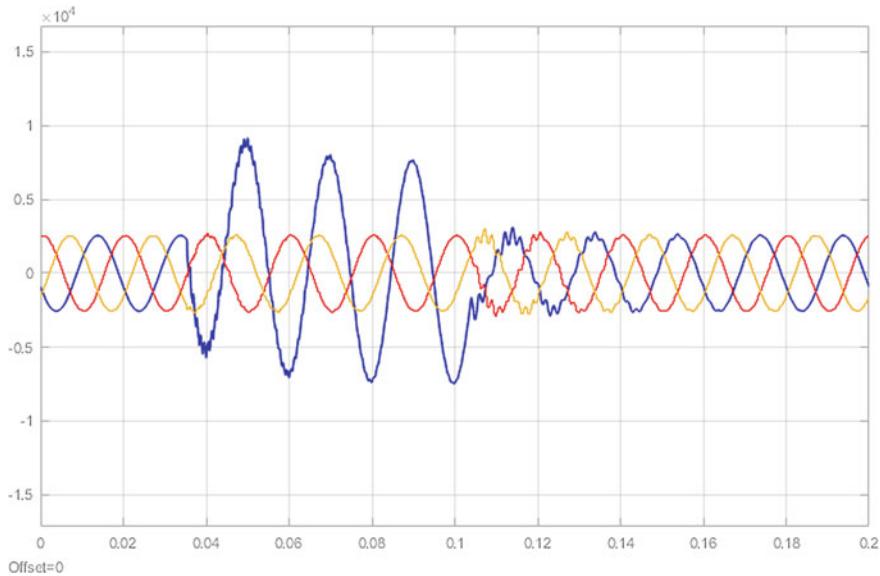
The circuit model diagram established in MATLAB/SIMULINK software is shown in Fig. 3.



**Fig. 2** Flow chart of traveling wave fault location



**Fig. 3** Simulation model diagram



**Fig. 4** Single-phase short-circuit power supply side three-phase voltage waveform

The blue line in the following simulation diagram represents phase A, the red line represents phase B, and the yellow line represents phase C. Take A-phase grounding short-circuit fault as an example to simulate and study single-phase grounding short-circuit. Firstly, the metal grounding fault of phase A is simulated, and the parameters of the three-phase variable fault circuit breaker in the simulation model are set. When  $t = 0.035$  s, a phase metal property grounding fault occurs, and when  $t = 0.1$ , the fault is removed. The simulated current is shown in Fig. 4 below.

It can be concluded from Fig. 4 that the three-phase current waveform at the M terminal presents a steady sine wave state in the steady state. At  $t = 0.035$ , a phase A short-circuit fault occurs, and the phase A current increases significantly. The resulting traveling wave diagram is shown in Fig. 5.

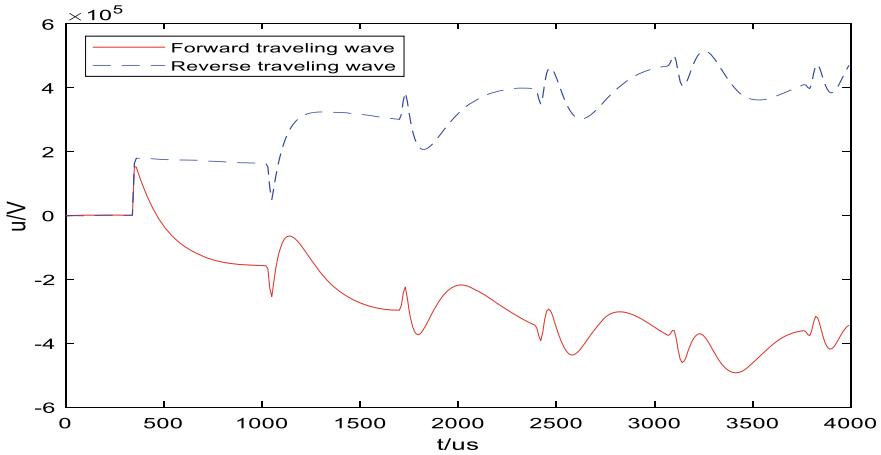
After phase mode transformation, three moduli of 0 mode,  $\alpha$  mode and  $\beta$  mode are obtained. The d8 wavelet transform is used to analyze the wavelet transform  $\alpha$ , and the detailed coefficients are shown in Fig. 6.

We can get  $t_1 = 36 \times 10^{-5}$  s,  $t_2 = 104 \times 10^{-5}$  s, and the set wave speed is  $v = 293, 290.5$  km/s.

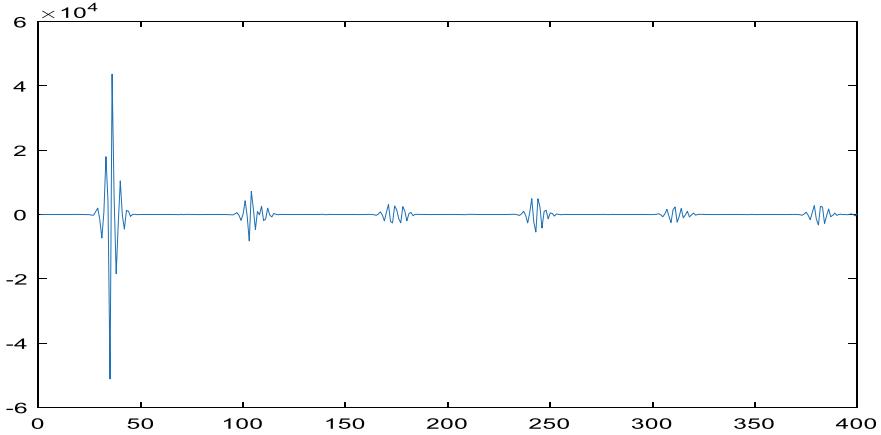
The measured distance is

$$x = \frac{v \times (t_2 - t_1)}{2} \approx 99.719 \text{ km} \quad (6)$$

The error distance is 0.281 km, and the error percentage is 0.28%, which meets the ranging requirements.



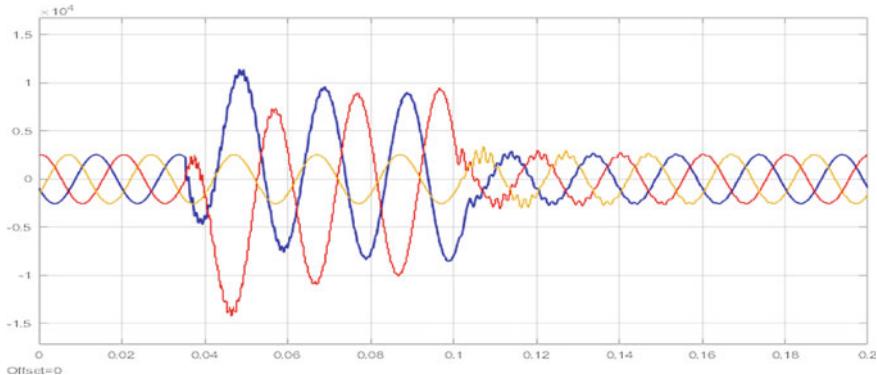
**Fig. 5** Single-phase short-circuit traveling wave diagram



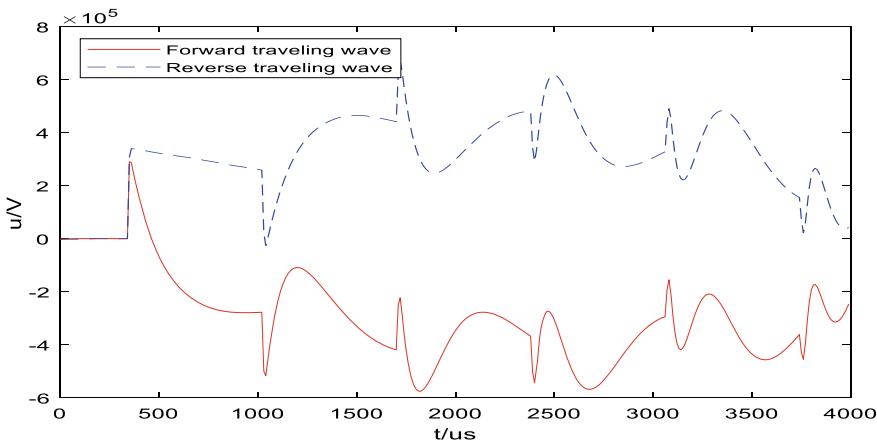
**Fig. 6** Single-phase short-circuit wavelet analysis diagram

Take the A and B phase grounding short-circuit faults as an example to simulate the single-phase grounding short-circuit. Firstly, the A-phase grounding fault is simulated, and the parameters of the three-phase variable fault circuit breaker in the simulation model are set. When  $t = 0.035$  s, a phase metal property grounding fault occurs, and when  $t = 0.1$ , the fault is removed. The simulated current is shown in Fig. 7.

It can be concluded from Fig. 7 that the three-phase current waveform at the M terminal presents a steady sine wave state in the steady state. At  $t = 0.035$ , a short-circuit fault occurs in the A and B phases, and the A and B phase currents increase significantly. The resulting traveling wave diagram is shown in Fig. 8.



**Fig. 7** Waveform of two-phase short-circuit current



**Fig. 8** Waveform of two-phase short-circuit current

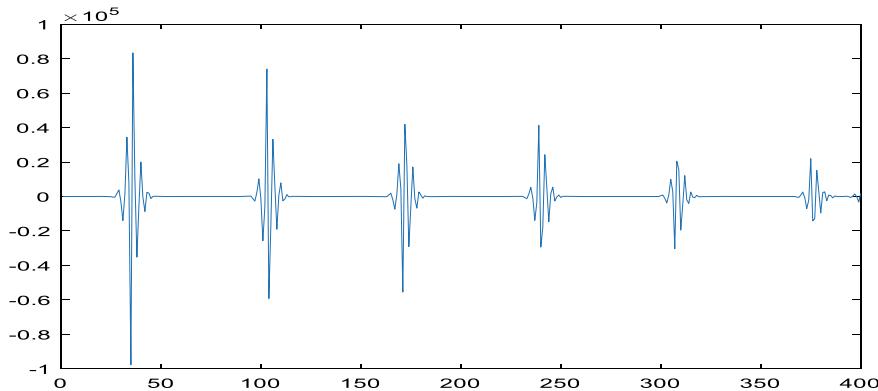
After phase mode transformation, three moduli of 0 mode,  $\alpha$  mode and  $\beta$  mode are obtained. The d8 wavelet transform is used to analyze the wavelet transform  $\alpha$ , and the detailed coefficients are shown in Fig. 9.

We can get  $t_1 = 36 \times 10^{-5}$  s,  $t_2 = 103 \times 10^{-5}$  s, and the set wave speed is  $v = 293, 290.5$  km/s.

The measured distance is

$$x = \frac{v \times (t_2 - t_1)}{2} \approx 98.252 \text{ km} \quad (7)$$

The error distance is 1.748 km, and the error percentage is 1.74%, which meets the ranging requirements.



**Fig. 9** Two-phase short-circuit wavelet analysis diagram

## 4 Conclusion

The paper mainly uses wavelet transform as a mathematical tool to analyze and process the fault information of the transmission line, and puts forward the theoretical method and distance measurement on the basis of traveling wave theory. This useful information can be extracted and analyzed by wavelet transform. After the phase selection of the fault data is completed by wavelet transform, the traveling wave theory is used for fault location. The error distance is 0.281 km in the 100 km single-phase short-circuit location experiment. The error percentage is 0.28%. In the 100 km two-phase short-circuit ranging experiment, the error distance is 1.748 km, and the error percentage is 1.74%. It verifies the feasibility of traveling wave ranging and obtains good experimental results.

**Acknowledgements** This research was funded by the National Natural Science Foundation of China (grant number 61672226, 61903137).

## References

1. Korkali, M.: Traveling-wave-based fault-location technique for transmission grids via wide-area synchronized voltage measurements. *IEEE Trans. Power Syst.* **27**(2), 1003–1011 (2012)
2. Qing, D., Yuan, Z., Zhigang, L.: A locating method of earth faults in large-scale power grid by using wide area measurement system. *Proc. Csee* **33**(31), 140–146 (2013)
3. Guangbin, Z., Hongchun, S., Jilai, Y.: Optimal placement of traveling wave current fault location devices in 220 kV power grid. *Proc. Chin. Soc. Electr. Eng.* **34**(34), 6246–6253 (2014)
4. Rui, L., Cheng, X., Fei, W.: Optimal Deployment of Fault Location Devices Based on Wide Area Travelling Wave Information in Complex Power Grid. *Transactions of China Electrotechnical Society* (2016)
5. Arifin, F.M., Hasan, M., Mahyudin, I., et al.: Development of fault distance locator for underground cable detection. *J. Phys. Conf. Ser.* **1432**, 012014 (2020)

6. Huang, R., Li, X., et al.: Research on combined traveling wave fault location of overhead line-cable hybrid line and influencing factors. *Power Syst. Prot. Control* **46**(5), 73–81 (2018)
7. Zhang, W., Xiao, X., Zhou, K., et al.: Multi-cycle incipient fault detection and location for medium voltage underground cable. *IEEE Trans. Power Deliv.* 1–1 (2016)
8. Kwon, G., Lee, C.K., Lee, G.S., et al.: Off-line fault localization technique on HVDC submarine cable via time-frequency domain reflectometry. *IEEE Trans. Power Deliv.* **32**(3), 1626–1635 (2017)
9. Gilany, M., Ibrahim, D.K., Eldin, E.: Traveling-wave-based fault-location scheme for multiend-aged underground cable system. *IEEE Trans. Power Deliv.* **22**(1), 82–89 (2006)
10. Li, Y., Wu, L., Li, J., et al.: DC fault detection in MTDC systems based on transient high frequency of current. *IEEE Trans. Power Deliv.* **34**(3), 950–962 (2019)
11. Huang, Q., Zhen, W., Pong, P.W.T.: A novel approach for fault location of overhead transmission line with noncontact magnetic-field measurement. *IEEE Trans. Power Deliv.* **27**(3), 1–1 (2013)
12. Franca, R.L., Junior, F.S., Honorato, T., et al.: Traveling wave-based transmission line earth fault distance protection. *IEEE Trans. Power Deliv.* **36**(2), 544–553 (2020)

# PCNetOP: Partial Completion Network with Order Prediction



Yifan Wang and Yongping Xie

**Abstract** Self-supervised methods can solve scene de-occlusion without a modal and order annotation, while supervised methods can only parse visible parts, leading to incomplete and unstructured scenario interpretation. The existing supervised methods that try to solve scene de-occlusion also need many manual annotations of invisible masks, which are costly and inaccurate. PCNets that use self-supervised method are introduced to solve this problem through new and unified frameworks that restore the hidden scene structure without the need for ordering and a modal annotations. This is achieved by Partial Completion Network (PCNet) -mask (M) and -content (C), which restore the components of the object mask and content respectively in a self-supervised way. We find that the accuracy of its order recovery in complex scenes is still low. In response to this problem, we propose a new occlusion order recovery method called PCNetOP (Partial Completion Network with order prediction), which combines the attention mechanism to redesign the network structure and training strategy to improve the occlusion complement effect in complex scenarios. Experiments show that our method solves the problem of self-supervised occlusion completion better than PCNet-M.

**Keywords** De-occlusion · Self-supervised learning · Semantic segmentation

## 1 Introduction

A scene understanding system should be able to deal with modal perception, that is, the region directly visible to perception, and the complete structure of perception entities, including invisible parts. The emergence of advanced deep networks and large-scale annotation datasets facilitates many scene understanding tasks, such as object detection [1–4], scene parsing [5–7], and instance segmentation [8–11].

Nonetheless, these tasks focus mainly on modal perception, and amodal perception remains under-explored so far. One of the key problems in modal perception is

---

Y. Wang · Y. Xie ()

School of Information and Communication Engineering, Dalian University of Technology,  
Dalian, Liaoning, China  
e-mail: [xieyp@dlut.edu.cn](mailto:xieyp@dlut.edu.cn)

the de-occlusion of the scene, which consists of restoring the latent occlusion order and fulfilling the sub-task of the invisible part of the occluded object.

To solve scene de-occlusion, one possible way is to train a model that can predict the occlusion order and the amodal mask (i.e., the full instance mask). There are some ways to get such true values from synthetic data [12, 13] or from manual annotation on real data [14–16], but each method has specific limitations. The former brings an ineluctable domain gap between the training data and the testing scene. The latter depends on different understandings among annotators to divide the obscured boundaries, thus it might create bias. This method requires repeated annotations from different annotators to reduce noise, and therefore the method is laborious and expensive. A more practical and extensible method is to get information from the data itself for scene de-occlusion rather than annotations.

In [17], the author addresses this problem in a self-supervised way, which can handle scene de-occlusion on real data without a manually annotated occlusion order or amodal mask. Without a real value, the end-to-end supervised learning framework no longer works. Thus, the author introduces a unique notion of partial completion of blocked objects. However, in terms of order recovery, the proposed method is based on the secondary inference of amodal completion results as part of the post-processing of amodal completion, which actually performs poorly on COCOA datasets. We present a new network structure based on this problem, directly predicting the occlusion order during the model inference phase. And we also add a light-weight attention mechanism applicable to this network to achieve a better result.

## 2 Related Work

### 2.1 Amodal Mask Completion

In the unsupervised domain, some works try to use depth to estimate whether the target is occluded by other objects. However, depth is unreliable in occlusion reasoning. The assumption proposed by these works that farther objects are blocked by close objects is not always true. In the field of supervision, some works manually annotate occlusion order [14, 15] or rely on synthetic data [13] to learn sorting in a fully supervised manner.

Modal segmentation, such as semantic segmentation [6, 7] and instance segmentation [8–10], aims to assign classification or object labels to visible pixels. Existing modal segmentation methods cannot solve the problem of de-occlusion. Unlike modal segmentation, the purpose of amodal instance segmentation is to detect objects as well as to recover their complete masks. Other works use a fully supervised learning approach with manual annotation [14–16] or synthetic data [13]. As mentioned above, manual annotation of invisible masks is costly and inaccurate. Methods that rely on synthetic data also face domain gaps. Conversely, the self-supervised way converts

the modal mask into an amodal mask. This unique ability takes the challenge out of the training of amodal instance segmentation networks without the need for manual amodal annotations.

## 2.2 Multi-task Learning

Compared with training separate models, multi-task learning aims to improve the learning efficiency and prediction accuracy in each task [18]. It can be considered as a method of inductive knowledge transfer to improve generalization ability by sharing domain information between complementary tasks.

In computer vision, there are many works using multi-task learning methods. Many people focus on semantic tasks such as classification and segmentation or classification and object detection.

DeepMask [19] uses CNNs to generate segmentation proposals instead of bounding box proposal algorithms with less information, such as selection search, MCG, etc. After the feature extraction, the feature map is inputted in the two brother branches. The top branch is based on the CNN object suggestion method, predicting a class-unknown segmentation mask, and the bottom branch marks the possibility that the estimated patch is centered on the whole object. The two branches share the same parameters of the network.

SharpMask [20] contains a bottom-up feed-forward network and a top-down network. The former is for generating coarse semantic segmentation masks, and the latter refines these masks using refinement modules. The bottom-up CNN architecture in SharpMask produces a thick mask encoding. The exported mask encoding is then fed into the top-down structure, where an optimization module un-pools the mask using matching features from the bottom-up module. This process runs on until reintegrating the final object mask.

## 2.3 Attention Mechanisms

Attention mechanisms have been proved to be helpful for various computer vision tasks such as image classification and image segmentation. A successful example of this is SENet [21], which can simply squeeze each 2D feature graph to efficiently construct the interdependencies between channels. CBAM [22] further advances this idea by introducing spatial information coding through convolution with large-scale kernels.

Non-local/self-attention networks have recently been very popular due to their ability to establish spatial or channel attention. But unlike these approaches that exploit costly and cumbersome non-local or self-attention blocks, CA [23] considers a more efficient capture of location information and channel relationships to enhance the feature representation of mobile networks, and performs much better than other

attention methods (e.g., SENet and CBAM) with the lightweight property by decomposing two-dimensional global pool operations into two one-dimensional coding processes.

## 3 Methods

### 3.1 Summary

PCNets are the networks that use a self-supervised method to solve the problem of scene de-occlusion, which include two parts: PCNet-M and PCNet-C. PCNet-M processes the mask completing part, and PCNet-C processes the content completing part. The model has been tested on COCOA and KINS datasets respectively. The experiments show that the model's accuracy on the COCOA dataset is much lower than that on the KINS dataset.

After analyzing the above problems, we think that this is due to the scene-oriented differences between the two datasets. The COCOA dataset is more close-shot than the KINS data set for road street view, in which indoor scenes are the majority. Thus the mask of the COCOA dataset is inevitably more complex and the geometry is more irregular, which negatively affects the accuracy of the model. Moreover, PCNets depend on the occlusion order prediction, and the order prediction results are got from the quadratic prediction of imperfect PCNet-M prediction results. As a result, the order prediction results are lower, which makes the final mask complement results worse.

After the above analysis, we solve the problem from two aspects:

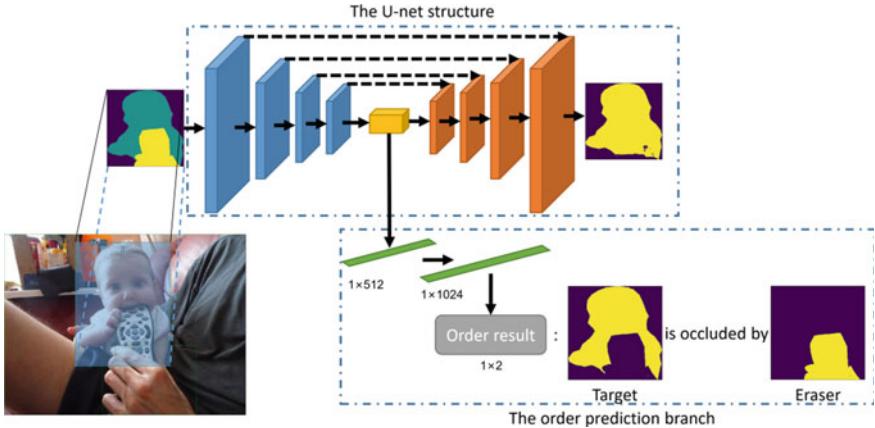
We propose a network that can directly predict the occlusion order while completing the occlusion mask segmentation. The training label is predicted by examining the intersection area of the occlusion mask truth value and the eraser. We design a new loss and combine the attention mechanism to optimize the model results.

We introduce an attention mechanism to guide the model to filter irrelevant features, enhance informative features and extract the more informative features.

### 3.2 Rehabilitation of Occlusion Networks

#### 3.2.1 Network Structure

According to PCNets, for the complex occlusion scene, the object with higher-order occlusion order may indirectly block the target instance, so it is necessary to recover the occlusion order. The eraser as model input at the inference stage shall consist of the eraser of all occlusion order classes, i.e., Formula (1) shows



**Fig. 1** PCNetOP network structure

$$\text{eraser}_{\text{all}} = \text{eraser}_1 \cup \text{eraser}_2 \cup \dots \cup \text{eraser}_n \quad (1)$$

Besides, for the task of unlabelled occlusion order recovery, including PCNets, previous works adopt the method of judging occlusion order by using mask complement results. However, mask complement results are not accurate and therefore this prediction method can lead to a higher misjudgement rate in the COCOA dataset than in the KINS dataset. We consider this problem as a quadratic prediction problem, that is, using the prediction results to predict another task.

For this secondary prediction problem, we can add an additional task branch to train together based on segmentation tasks, so as to predict the occlusion order directly in the model reasoning stage. In works focusing on multitask training, such as DeepMask and SharpMask, the authors add a predictive branch for their segmentation task. SharpMask is more concise and efficient than DeepMask in terms of the addition of the branch, and the model is closer to the U-Net structure. Therefore, we add the order prediction branch adapted to our task based on the U-Net network structure, with reference to SharpMask.

The model structure we design is shown in Fig. 1. We modify the U-Net structure to add a  $1 \times 1$  convolution to its down sampling end and output the occlusion sequence prediction results through three fully connected layers. Since the adopted loss is Cross-Entropy Loss, the output size is  $1 \times 2$ . We name the network PCNetOP (PCNet with order prediction).

### 3.2.2 Attention Mechanism

In addition to the impact of the accuracy of the occlusion order prediction, another reason for the differences in the Mask completion results is the model's insufficient

grasp of the global information. Native U-Net cannot accurately capture the connection between the Target and eraser channels due to its lack of global information retrieval.

The CA module, through introduction of the global pooling of H and W orientations, gets the global information that general pooling cannot extract, thus improving global information extraction by the SE attention mechanism. The CBAM attention mechanism is divided into two modules: ChannelAttention and SpatialAttention. ChannelAttention performs feature enhancement by mean and maximum pooling stacking, and SpatialAttention can average the entire tensor to enhance the edge information, and enhance the acquisition of shape information. We add the CA + SA attention mechanism to each down block as well as the in Conv block. We replace the channel attention mechanism in the CBAM with the CA module, and connect it with the spatial attention mechanism to constitute a new attention mechanism, thus further improving our model performance.

### 3.3 Loss Function Design

We determine the true occlusion order by examining the relationship between the mask ground truth and the eraser, as shown in Eq. (2). When the intersection is not equal to 0, we set 0 to express the occlusion case, and when the intersection is equal to 0, we set 1 to express the case of no occlusion.

$$\text{OCCvalue} = \begin{cases} 1 & \text{Target} \cap \text{Mask} = 0 \\ 0 & \text{Target} \cap \text{Mask} \neq 0 \end{cases} \quad (2)$$

The Cross-Entropy Loss is calculated using the  $1 \times 2$  size output of the model for training. The training loss calculation formula is shown in Formula (3), where  $P_{\theta\text{order}}^{(m)}$  is the model order branch output.

$$L_{\text{order}} = \sum_{A, B \in D} L(P_{\theta\text{order}}^{(m)}, \text{OCC value}) \quad (3)$$

When training with the occlusion mask complement branch, the multi-task learning weight setting problem is involved. And we find that when SharpMask is added directly, the weight setting is more difficult, and the task converges more easily to the occlusion order branch task.

In order to solve this problem, we adopt the method of adaptive weight adjustment. By giving a loss weight training gradient, AutomaticWeightedLoss dynamically adjusts the weight value during training. We employ this method for two losses to mitigate the trend that the dual task converts to the single task direction. Finally, the total loss function is shown in Formula (4) as follows:

$$L_{\text{sum}} = \frac{1}{2\sigma_1^2} L_{\text{Mask}} + \frac{1}{2\sigma_2^2} L_{\text{Order}} + \log(1 + \sigma_1^2) + \log(1 + \sigma_2^2) \quad (4)$$

## 4 Experiment

### 4.1 Experimental Details

Our experiment sets the input image size as  $256 \times 256$ , eraser\_front\_prob as 0.5 to train the classification branch better. We use the Adam optimizer (batch 8, val\_batch\_size 32, lr 1e-3, iterations 62 k). We use GTX2080 experimental equipment. The training framework is PyTorch.

The COCOA dataset is a subset of the COCO2014, at the same time labelled into two pairs of order, modal mask, and amodal mask. We train our model on this dataset and compare the result with PCNet-M. Figure 2 shows the loss drop curve.

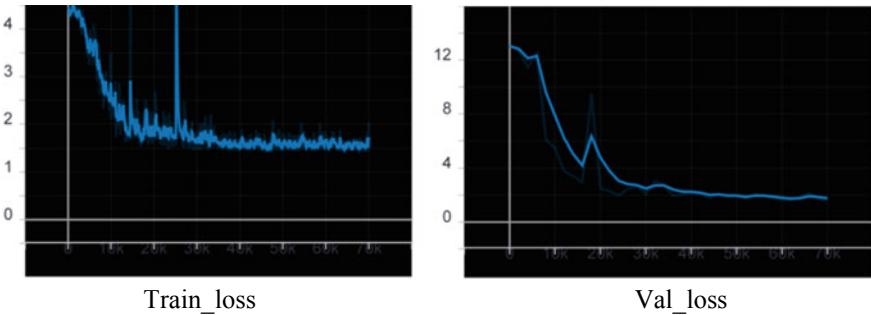
### 4.2 Results

To test the new occlusion order prediction method, we also modify the model output in the inference section. Model output is a tensor of size  $1 \times 2$ . First, the output of the model order branch gets through a softmax. Then, according to the first element of the tensor, we use 0.5 as the threshold to determine the category, assigning 1 and -1 respectively. Consistent with PCNet, a pair of non-adjacent object occlusion categories are assigned to 0. And when generating order\_matrix, the mask of each pair of targets is traversed once, and then the final order\_matrix is obtained by taking the inverse number with a positive diagonal line as the axis. We do amodal mask completion with this order\_matrix. We test scores on COCOA datasets, and Table 1 shows the comparison results with the original method.

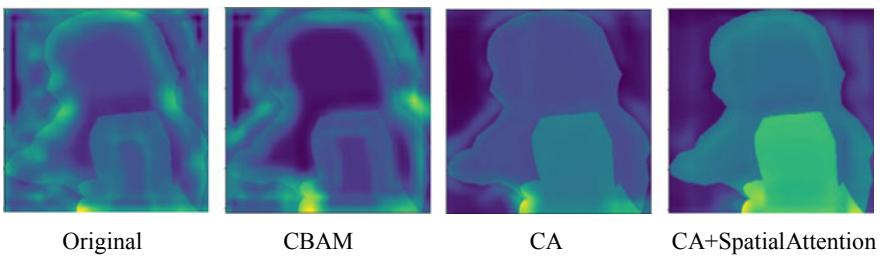
Furthermore, for the attention mechanism part, we take the middle layer output of the fourth down block in the network for visualization to observe the effect of the attention mechanism on the performance of the model. According to Formula (5), all channels are added to the average value and then superimposed with the original image:

**Table 1** Comparison with PCNet-M original method on COCOA dataset

Method	acc_allpair	acc_occpair	mIoU	pAcc
PCNet-M	-	0.871	0.8135	-
PCNetOP	0.96965	0.91632	0.81781	0.86752



**Fig. 2** Training loss curve



**Fig. 3** Visual comparison of different attentions

$$\text{superimposed\_img} = \text{feature\_resized} \times 10 + \text{img} \quad (5)$$

where  $\text{img}$  is the superposition of the input two channels of normalized modal mask and the eraser, i.e., the Formula (6) shows.

$$\text{modal\_mask}[\text{eraser} = 1] = 2 \quad (6)$$

Figure 3 is the visual comparison of intermediate results. Compared with CBAM, in the CA module, the filtering effect of irrelevant features is very significant. But the CA module lacks the spatial attention mechanism to extract spatial information. Therefore, we replace the channel attention module in CBAM with the CA module for this task.

### 4.3 Ablation Experiments

In this part, we show the influence of different optimization strategies on the experimental results.

**Table 2** Attention mechanisms

Method	acc_allpair	acc_occpair	mIoU	pAcc
Original model	0.95065	0.82706	0.76812	0.82545
CBAM	0.95419	0.84416	0.79842	0.87186
To each block	0.95878	0.86486	0.81123	0.87761
CA module	0.9568	0.85667	0.80568	0.87634
CA + SpatialAttention	0.95781	0.86126	0.80741	0.87547
ReLU6 replaced with ReLU	0.96124	0.8769	0.81533	0.87501

**Table 3** Loss design

Method	acc_allpair	acc_occpair	mIoU	pAcc
Original modal	0.95065	0.82706	0.76812	0.82545
New order	0.97207	0.92748	0.80673	0.86315
With autoweightloss	0.96965	0.91632	0.81781	0.86752

Table 2 shows the effect of the attention mechanism on model performance. The results show that it is better to integrate attention modules into down blocks than to add only one layer. The combination of the CA module and SpatialAttention is better than a single CA module. When we replace ReLU6 with the ReLU activation function, the new attention mechanism exceeds the CBAM performance.

Table 3 shows the results of adding occlusion prediction branches. As the result shows, the model training results can be improved by using the AutomaticWeightedLoss.

## 5 Conclusion

We improve the results of the problem of poor occlusion order prediction in self-supervised de-occlusion. We provide a recipe to deliver the occlusion order straightforwardly by adding the occlusion prediction branch and redesigning the occlusion order label. The proposed method can effectively improve the occlusion order prediction performance and further improve the mask completion results.

Furthermore, we argue that our approach can be applied to solve mutual occlusion problems by making corresponding modifications to the post-processing section. As the COCOA dataset contains only unilateral occlusion dimensions, performance on mutual occlusion problems cannot be evaluated. We argue that the mutual occlusion problem can be better solved by modifying the modal mask and eraser generation methods, and by predicting both occlusion orders of unilateral occlusion and mutual occlusion.

## References

1. Pedro F.F., Ross B.G., David M.: Cascade object detection with deformable part models. In: CVPR, pp. 2241–2248. IEEE (2010)
2. Shaoqing, R., Kaiming, H., Ross, G., et al.: Faster r-cnn: towards real-time object detection with region proposal networks. In: NIPS (2015)
3. Jiaqi, W., Kai, C., Shu, Y., et al.: Region proposal by guided anchoring. In: CVPR (2019)
4. Kai, C., Jiaqi, W., Shuo, Y., et al.: Optimizing video object detection via a scale-time lattice. In: CVPR (2018)
5. Ziwei, L., Xiaoxiao, L., Ping, L., et al.: Semantic image segmentation via deep parsing network. In: ICCV (2015)
6. Liang-Chieh, C., George, P., Iasonas, K., et al.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4), 834–848 (2017)
7. Hengshuang, Z., Jianping, S., Xiaojuan, Q., Xiaogang, W., Jiaya, J.: Pyramid scene parsing network. In: CVPR, pp. 2881–2890 (2017)
8. Jifeng, D., Kaiming, H., Yi, L., et al.: Instance-sensitive fully convolutional networks. In: ECCV, pp. 534–549. Springer (2016)
9. Kaiming, H., Georgia, G., Piotr, D., et al.: Mask r-cnn. In: ICCV (2017)
10. Kai, C., Jiangmiao, P., Jiaqi, W., Yu, X., et al.: Hybrid task cascade for instance segmentation. In: CVPR, pp. 4974–4983 (2019)
11. Jiaqi, W., Kai, C., Rui, X., et al.: Carafe: content-aware reassembly of features. In: ICCV (2019)
12. Kiana, E., Roozbeh, M., Ali, F.: Segan: segmenting and generating the invisible. In: CVPR, pp. 6144–6153 (2018)
13. Yuan-Ting, H., Hong-Shuo, C., Kexin, H., et al.: Sail-vos: semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In: CVPR, pp. 3105–3115 (2019)
14. Yan, Z., Yuandong, T., Dimitris, M., Piotr, D.: Semantic amodal segmentation. In: CVPR, pp. 1464–1472 (2017)
15. Lu, Q., Li, J., Shu, L., et al.: Amodal instance segmentation with kins dataset. In: CVPR, pp. 3014–3023 (2019)
16. Follmann, P., Knig, R., Hrtinger, P., et al.: Learning to See the Invisible: End-to-End Trainable Amodal Instance Segmentation (2018)
17. Xiaohang, Z., Xingang, P., Bo, D., et al.: Self-supervised scene de-occlusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
18. Alex, K., Yarin, G., Roberto, C.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. Pedro, O.P., Tsung-Yi, L., Ronan, C., et al.: Learning to refine object segments. In: European Conference on Computer Vision (2016)
20. Tianshui, C., Liang, L., Xian, W., et al.: Learning to Segment Object Candidates via Recursive Neural Networks (2016)
21. Jie, H., Li, S., Gang, S.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
22. Sanghyun, W., Jongchan, P., Joon-Young, L., et al.: CBAM: Convolutional Block Attention Module. Springer, Cham (2018)
23. Qibin, H., Daquan, Z., Jiashi, F.: Coordinate Attention for Efficient Mobile Network Design (2021)

# Fastener Identification Method Based on Two-Stage Positioning



Yan Li , Hongbin Liu , and Zhigang Liu

**Abstract** With the rapid development of railroad transportation in recent years, the traffic safety and maintenance decisions put forward more and more stringent requirements. Due to the harsh environment and other factors, railroad fasteners often break or miss. This poses a threat to railroad safety. Therefore, railroad fastener status detection is one of the important means to ensure the safe operation of the railroad, and the identification of railroad fasteners is the important technology of railroad fastener status identification. In the paper, a fastener identification method based on two-stage positioning is proposed. In the first step, the paper identifies the steel rails and rail sleepers, and then roughly locates the fasteners according to the relative position relationship of the steel rails, rail sleepers and fasteners. In the second step, our method integrates the area statistics algorithm and the template matching algorithm to achieve the precise positioning of fasteners. The experimental results show that our proposed method can effectively identify fastener.

**Keywords** Fasteners · Identification · Two-stage · Positioning

## 1 Introduction

Railway is a form of transportation in China. It is the backbone of China's transportation system. In recent years, China's comprehensive national power has surged in pace with its increasingly boomed economy. Therefore, Railroad transportation has also developed speedily. At the same time, the issue of railroad transportation safety has gained the attention of scholars, enterprises and research institutions.

In the past time, universities like Beijing Jiaotong University and Southwest Jiaotong University and Institutes like China Academy of Railway Sciences gradually

---

Y. Li · H. Liu ()

Shandong Key Laboratory of Intelligent Buildings Technology, School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China  
e-mail: [liuhongbin19@sdu.edu.cn](mailto:liuhongbin19@sdu.edu.cn)

Z. Liu

School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China

researched rail detection techniques based on image processing technology. Hong Fan [1] designs an adaptive crossover method to locate the fastener area. The algorithm has strong robustness. But the crossover method does not identify the fastener region precisely. Therefore, Hong Fan [1] intercepts part of the images of fasteners as templates to perform template matching in the region where the fasteners were coarsely positioned. Qing Song [2] proposes a high-speed railway fastener detection and positioning algorithm based on convolutional neural network. Fan Xu [3] proposes a fastener localization method based on deep learning. The method combines the template matching and the deep learning to use different fastener recognition algorithms according to different track scenes. Hong Fan [4] uses the template matching to identify fasteners. Yijin Qiu [5] proposes a fastener positioning algorithm based on double template matching (DTM). The double template contains the rail template and the fastener template. Xiukun Wei [6] proposes a fastener location algorithm based on projection analysis and DB4 wavelet transform. Jianwei Liu [7] and Zhiyong Peng [8] propose an identification method for the hexagon nut in the fastener center. The precise positioning of fasteners can eventually be achieved. Ruxun Xu [9] proposes a method of fastener recognition based on multi feature fusion and training.

The crossover method has short running time, but the accuracy of fastener region identification is not high. The template matching, furthermore, can improve the accuracy of identifying fasteners, but the running speed is not high.

In view of the above problems, this paper proposes a fastener identification method based on two-stage positioning. First, the rail and rail sleeper are identified, and then the fasteners are roughly positioned according to the relative positions of the steel rail, the rail sleeper and the fasteners. Then, the result obtained from coarse localization is binarized and processed morphologically. The edges of the morphologically processed result are detected and regional statistics are performed. Since the edges detected by the fastener integrity and fastener incomplete cases are different, this paper sets a threshold  $T$  by experiment, and the maximum obtained from the regional statistics is used as the index to judge whether the fastener is complete or not. If the fastener is complete, the position of the maximum is used directly as the fastener identification result. If the fasteners are incomplete, the method of template matching is used to identify the fasteners. This is because the method of template matching is more accurate in the case of incomplete fasteners. This method improves the accuracy of fastener identification and also accelerates the operation speed.

This paper is organized as follows. In the first section, the background and significance of this paper are introduced, and the current status of domestic research and the existing problems are described. In Sect. 2, the process of the fastener identification method based on two-stage positioning proposed in this paper is described. In Sect. 3, the experimental platform and data of this paper are introduced. Finally, some conclusions are given in Sect. 4.

## 2 The Fastener Identification Method Based on Two-Stage Positioning

### 2.1 Initial Identification of the Fasteners

This paper first identifies the rail sleeper and the steel rail. As the installation of fasteners follows certain guidelines, the position relationship between the fasteners and the rail sleeper and the steel rail remains unchanged. Therefore, through the relative position between the three to achieve the initial positioning of the fasteners. As a result, the identification range is narrowed.

**Positioning of the rail sleeper.** It can be observed that the edge of the rail sleeper shows a horizontal long straight line at the edge, and the upper edge and the lower edge are parallel. Conversely, the straight lines of the edges of the railway ballast are short and messy. The method of rail sleeper identification based on the Line Segment Detection(LSD) [10] algorithm is proposed by the paper [11]. According to the above characteristics, the LSD algorithm is used in the paper to detect the straight lines in the image. Then the detected lines are filtered. Finally, the recognition of the rail sleeper is realized. The overall algorithm steps are shown as Table 1.

**Positioning of the steel rail.** Analyzing the railroad images, we can conclude that the steel rail is vertically distributed. Due to the friction of the vehicles, the brightness of the surface part of the steel rail is high compared to the surrounding objects. The rail recognition algorithm is proposed by the paper [12]. For this reason, the rail is recognized by the area statistical method shown in Table 2.

$$S_g(b) = \sum_{k=1}^{W_R} g(b+k) \quad (1)$$

where,  $1 \leq b \leq W_0 + W_R$ ,  $H_0$ .  $W_0$  is the height and width of the input image, respectively.

**Table 1** The method of rail sleeper identification based on LSD algorithm

---

Input image: acquired image of the railway line
Output image: localization image after identifying the rail sleeper
1. Detect the straight line of image $I$ with LSD algorithm and get the result lines
2. Sort out the long horizontal lines from the lines that meet the following conditions: length > threshold, and angle less than angle_thres
3. Calculate the distance between any two long horizontal lines and sort out the two (approximate) parallel lines whose distance is in the range of dis_thres1-dis_thres2
4. Calculate the medians of the two groups of y-values for the more concentrated distribution of line segments. Take the median values as the coordinates of the edges of the rail sleeper
5. End

---

**Table 2** The method of the steel rail identification based on area statistics

Input image: Acquired image of the railway line
Output image: Localization image after identifying the steel rail
1. Gray-scale projection of image $I$ in the vertical direction. Projection value $g(x)$ is the average grayscale value of column $x$
2. Define the width of the shiny part of the rail as $W_R$ . Sum the projection results in steps of 1 for each $W_R$ as a unit, and record the projection value as $Sg$ . The expression of $Sg$ is shown in Eq. (1)
3. Find the maximum value of the $Sg$ and the position of the maximum value is the position of the steel rail
4. End

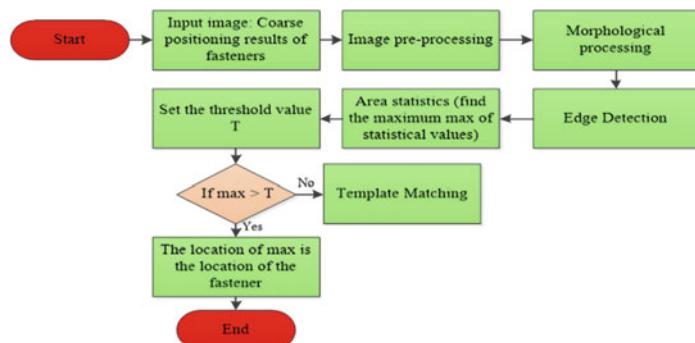
**Initial Positioning.** According to the relative position of the rail, the rail sleeper and the fastener, the fastener is positioned. Here are some assumptions. The coordinate of the upper edge of the rail sleeper is  $y1$ . The coordinate of the right edge from the steel rail is  $\times 1$ . The distance of the right fastener from the upper edge of the rail sleeper is  $d1$ . The distance of the right fastener from the right edge of the rail is  $d2$ . The width of the fastener is  $W$  and the length is  $L$ . Then the position coordinates of the right fastener in the horizontal direction are  $\times 1 + d2$ ,  $\times 1 + d2 + W$ , and the position coordinates of the right fastener in the vertical direction are  $y1 + d1$ ,  $y1 + d1 + L$ .

## 2.2 Accurate Identification of the Fasteners

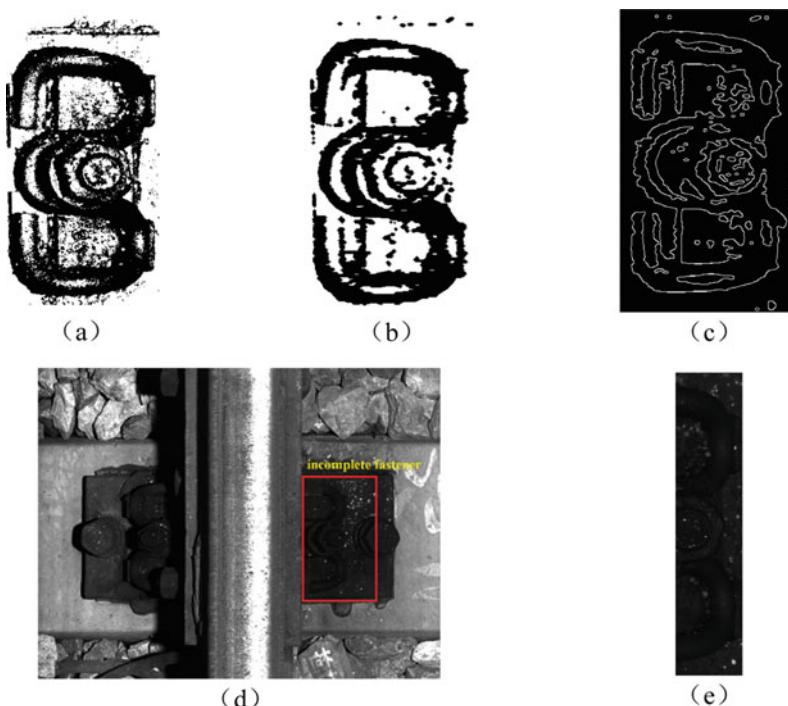
In the second stage of fastener localization, the result of coarse fastener localization is preprocessed. Then the preprocessed result is morphologically processed and detect edges. Subsequently, the area statistics of the edge detection result, max, are performed to obtain the maximum value of the area statistics. A threshold value  $T$  is set by a small number of experiments. The max and  $T$  are compared to determine whether the fastener is complete or not. Then select different methods to identify different cases. The specific algorithm flow is shown in Fig. 1.

**Image pre-processing.** The preliminary results are saved as the image to be measured. To make the image clearer, median filtering and histogram equalization are applied to the image to be measured. Then the enhanced image is binarized using the *imbw* function provided by MATLAB. Here, the *graythresh* function is used to find a suitable threshold for the image to be measured by the maximum interclass variance method, and then the result of the function is used as the threshold for the binarization process. The result of the binarization is shown in Fig. 2a.

**Morphological processing.** In this paper, we use the structure shown in Structure 1 to do the closed operation first and then do the open operation, and finally use the structure shown in Structure 2 to do the closed operation. The results are shown in Fig. 2b.



**Fig. 1** The flow chart of the algorithm for precise positioning of fasteners



**Fig. 2** **a** Result of banalization; **b** result of morphological processing; **c** result of edge detection; **d** the image of the incomplete fastener; **e** the template

$$\begin{array}{c} \left[ \begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \quad \left[ \begin{array}{cccc} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right] \\ \text{Structure 1} \qquad \qquad \qquad \text{Structure 2} \end{array}$$

**Edge detection.** Edge detection is performed on the results obtained from morphological processing, and the edge detection operator is the canny operator. This is because the canny operator is not easily disturbed by noise, the edge detection effect of the canny operator is a little better. The edge detection results are shown in Fig. 2c.

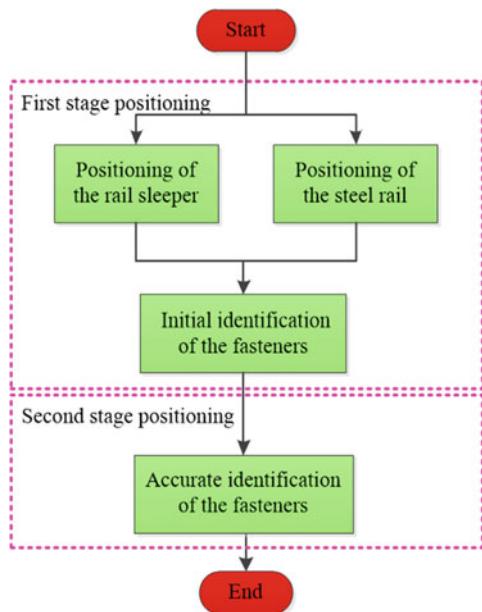
**Regional statistics.** First, the vertical projection is performed on the image  $I_1$  of edge detection. In other words, the pixel values of all pixel points with the same horizontal position in the graph are summed up. The horizontal projection of  $I_1$  is also performed. Assuming that the length of the fastener is  $L$  and the width is  $W$ . For the horizontal projection, the projection results are summed every  $L$  lengths in steps of 1, and finally the  $y$  coordinate corresponding to the maximum value of the summation result is sought. This  $y$ -coordinate is the position of the upper edge of the fastener. (Fasteners generally miss a part in the horizontal direction, and are basically intact in the vertical direction. Incomplete fastener is shown in the Fig. 2d. The area marked out by the red rectangular box in the figure is the incomplete fastener. So it is not necessary to determine whether the fasteners are complete or not at this time). For the vertical projection result, the summation is performed once for each  $W$  length in steps of 1. The final summation maximum, max, and its corresponding  $x$ -coordinate are found. Set a suitable threshold value  $T$  by experiment. If  $\text{max} > T$ , the fastener is judged to be complete and the  $x$ -coordinate corresponding to max is the position coordinate of the left edge of the fastener. If  $\text{max} < T$ , the fastener is judged to be incomplete and then the fastener is identified by the method of the template matching.

**Template Matching.** Since this step is to identify incomplete fasteners, the right half of the fastener is intercepted when the template is created, and the specific template is shown in the Fig. 2e. The HOG features of the template are calculated to obtain the HOG feature histogram. Create a window with the size of the template. Slide the window over the fastener coarse positioning result in steps of 2 until the entire fastener coarse positioning result is traversed. Calculate the similarity coefficient between the HOG feature histogram and the template for each window. The formula for calculating the similarity coefficient is shown in (2). The window with the smallest similarity coefficient is taken as the result of the precise positioning of the fastener.

$$d(X, Y) = x^2 = \sum_{i=1}^n \frac{(x_i - y_i)^2}{x_i + y_i} \quad (2)$$

Up to this point, this paper has realized the precise positioning of the fastener.

**Fig. 3** Flowchart of the fastener identification method based on two-stage positioning



### 2.3 Summary of this chapter

This chapter is divided into two subsections to introduce the fastener identification method based on two stages. The overall algorithm flowchart is shown in Fig. 3.

## 3 Experiment

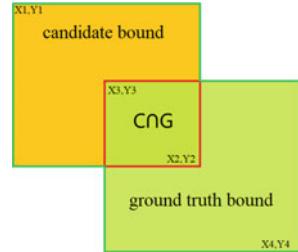
### 3.1 Fastener Identification Experimental Platform

The experimental process of this paper is all realized on the same platform, and the details of the fastener identification experimental platform are shown in Table 3.

**Table 3** Parameters of the experimental platform

CPU	Intel(R) Core(TM) i5-6200U CPU @2.30 GHz 2.40 GHz
Memory	4 GB
Operating system	Windows10
Exploitation environment	MATLAB R2016a
Programming language	MATLAB

**Fig. 4** Intersection and concatenation



### 3.2 Evaluation Index

In this paper, the intersection-over-Union (IOU) is used to measure the accuracy of the experimental results. The IOU, a concept used in target detection, is the ratio of the overlap of the resulting candidate bound to the original ground truth bound, i.e., the ratio of their intersection to the concatenation. The ideal case is complete overlap, i.e., a ratio of 1.

As shown in Fig. 4, the area marked by the red box is the intersection of the candidate box and the original marked box, denoted by  $C \cap G$ . The area marked in green is the concatenation of the candidate box and the original marked box, denoted by  $C \cup G$ .

The specific algorithm of IOU is shown in formula (3)

$$\text{IOU} = \frac{\text{area}(C \cap G)}{\text{area}(C \cup G)} \quad (3)$$

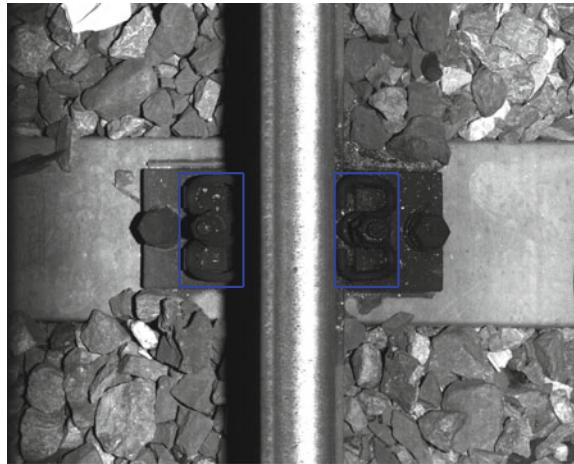
### 3.3 Experimental Results

**Experiments on fastener identification method based on two-stage positioning.** By applying the two-stage fastener identification algorithm based on Sect. 2, the fasteners are finally identified and marked out with blue boxes. The experimental results are shown in Fig. 5. The fastener area is marked by the blue rectangle in the Fig. 5.

### 3.4 Comparison Experiments

Template matching is the earliest and most widely used fastener recognition algorithm. In this paper, the fastener recognition algorithm based on template matching is used as a baseline method to compare with the algorithm proposed in the paper.

**Fig. 5** Result of fastener positioning



The fastener identification algorithm based on template matching is to firstly realize coarse identification of fasteners according to Sect. 2.1. Then the fastener template is produced. The similarity coefficient between the HOG features of the fastener template and the HOG features of the fastener coarse identification result is calculated. Finally, the accurate identification of fasteners is realized.

After the program was written and run in MATLAB software, the precise positioning of the fasteners was finally achieved. Then the IOU was calculated and the experimental results were counted. The experimental data of the algorithm and template matching algorithm in this paper are shown in Table 4.

From the final experimental results, it is shown that the accuracy of the method used in this paper has a little higher IOU, and its running time is faster than the template matching method.

## 4 Conclusion

In the actual detection system needs to achieve high real-time requirements. In order to solve the problem of long template matching operation time, this paper proposes a fastener identification method based on two stages. But the algorithm in this paper uses fewer samples. At the same time, since there is no sample of missing fasteners in the existing samples, the algorithm of this paper cannot realize the identification of missing fasteners when they are missing. At present, there is not much research related to fastener detection. But I believe that with the efforts of many enterprises, universities and research institutions, we will soon make breakthroughs in this field in this era of big data and the rapid development of machine vision field as well as machine learning field.

**Table 4** Compare experimental data

Data number	Fastener identification method based on two-stage positioning		The method of template matching	
	IOU (%)	Time (s)	IOU (%)	Time (s)
1	94.15	2.49	92.83	7.3760
2	97.80	2.34	95.66	7.5460
3	92.42	10.63	92.42	7.5470
4	97.12	2.37	95.97	7.1870
5	95	2.76	92.75	7.4060
6	90.99	2.61	95.09	7.7810
7	94.36	2.78	92.62	3.0150
8	95.95	2.74	94.78	2.9710
9	96	2.84	97.03	3.1560
10	93.68	2.19	89.26	2.6710
11	98.08	2.64	89.38	3.26
12	98.19	3.23	95.24	3.3140
13	96.2	3	89.16	15.1890
14	93.63	2.9	88.63	15.3910
15	93	2.89	83.96	15.6530
16	93.5	3.36	90.27	18.2650
17	96.39	3.7	94.09	18.03
18	92.81	3.44	94.12	18.1870
19	92.81	2.79	91.07	14.3270
20	91.8	2.22	83.36	14.2340
21	93.52	2.381	96.30	14.4190
Average	94.64	3.157	92.09	9.854

## References

1. Hong, F.: The Research of Algorithm of the Detection of Railroad Fastener Defects Based on Image. Southwest Jiaotong University (2012)
2. Qing, S., Yao, G., Jianan, J., Chun, L., Mengjie, H.: High-speed Railway Fastener Detection and Localization Method based on Convolutional Neural Network. arXiv (2019)
3. Fan, X.: The Research of Railroad Scene Recognition and Fastener Localization Method Based on Computer Vision. Beijing Jiaotong University (2018)
4. Hong, F., Pamela C. C., Yun, H., Bailin, L.: High-speed railway fastener detection based on a line local binary pattern. IEEE Signal Process. Lett. **25**(6), 788–792 (2018)
5. Yijin, Q., Xingjie, C., Zhaomin, L.: Rail fastener positioning based on double template matching. Complexity **2020**, 10 (2020)
6. Xiukun, W., Ziming, Y., Yuxin, L., Dehua, W., Limin, J., Yujie, L.: Railway track fastener defect detection based on image processing and deep learning techniques: a comparative study. Eng. Appl. Artif. Intell. **80**, 66–81 (2019)
7. Jianwei, L., Hongli, L., Xuefeng, N., Ziji, M., Chao, W., Shao, X.: A visual inspection system for accurate positioning of railway fastener. IEICE Trans. Inf. Syst. **103**(10), 2208–2215 (2020)

8. Zhiyong, P., Chao, W., Ziji, M., Hongli, L.: A multifeature hierarchical locating algorithm for hexagon nut of railway fasteners. *IEEE Trans. Instrum. Meas.* **69**(3), 693–699 (2020)
9. Ruxun, X., Jinyue, H., Wenzhe, Q., Jianjun, M., Hongqiang, Z.: Railway fastener image recognition method based on multi feature fusion. In IOP Conference Series: Materials Science and Engineering, vol. 397, No. 1, pp. 012119. IOP Publishing (2018)
10. Rafael, G., Jérémie, J., Jean-Miche, M., Gregory, R.: LSD: a line segment detector. *Image Process On Line* **2**, 35–55 (2012)
11. Feng, H., Jiang, Z., Xie, F., Yang, P., Shi, J., Chen, L.: Automatic fastener classification and defect detection in vision-based railway inspection systems. *IEEE Trans. Instrum. Meas.* **63**(4): 877–888
12. Shengwei, R., Qingyong, L., Guiyang, X., Qiang, H., Siwei, F.: The research of robust real-time steel rail surface abrasion detection algorithm. *China Railway Sci.* **32**(1), 25–29 (2011)

# EmbedLOF: A Network Embedding Based Intrusion Detection Method for Organized Attacks



Peng Chen , Yunfei Guo , Jianpeng Zhang , and Hongchao Hu 

**Abstract** To increase the detection rate of organized attacks in cyberspace, a new intrusion detection method, i.e., EmbedLOF, is proposed which combines the network embedding and outlier detection method. The proposed method first preprocesses the captured packets, generates network undirected graph, and calculates connected components of the undirected graph. Then the Embed algorithm is utilized to generate network embedding of each node for the connected components of the undirected graph. The network embedding uses low-dimensional vectors to represent latent features in network topology. Finally, the LOF algorithm is utilized to conduct the outlier detection for each node's embedding. Also, it issues alarms for possible intrusions. The experiment results show that the method obtains high recall scores and achieves more comprehensive and robust detections for organized attacks.

**Keywords** Network embedding · Outlier detection · Intrusion detection · Local outlier factor

## 1 Introduction

Since the rise of the Internet in the late twentieth century, cyberspace has increasingly become an indispensable space in people's daily work and life. However, due to the lack of awareness of cybersecurity among ordinary people and the defects in the design of cyberspace infrastructure itself, crimes in cyberspace are endless, and cyber attacks are difficult to eradicate. For instance, in China there were about 42.08 million IP addresses attacked by computer malicious programs in the first half of 2020, and there were about 220 high-rate DDoS (Distributed Denial of Service) attacks whose daily peak traffic exceed 10Gbps, according to the report of the National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC) in September 2020 [1]. These attacks have caused serious

---

P. Chen · Y. Guo · J. Zhang (✉) · H. Hu

Strategic Support Force Information Engineering University, Zhengzhou 450001, China

e-mail: [zjp@ndsc.com.cn](mailto:zjp@ndsc.com.cn)

losses to the interests of individuals, enterprises, and government departments. As a result, security issues in cyberspace have received widespread attention.

For defense against network attacks, or more broadly, network intrusions, the common practice of enterprises and government departments is to deploy an Intrusion Detection System (IDS). The machine learning based intrusion detection method is considered to be a promising detection method to deal with network intrusions, especially novel unknown attacks [2–6]. In [5], the authors proposed to use nonsymmetric deep autoencoder (NADE) for unsupervised feature learning, and proposed to use stacked NADE to form a deep learning classification model to solve the problems that traditional Network-based IDSs (NIDSs) require too much manual intervention, expertise, and have a high rate of error in detection results. Through experiments on real datasets, the authors proved that the proposed classification model had higher accuracy and recall. The authors in [6] proposed a combinatorial intrusion detection model (DRRS) based on deep recurrent neural network (DRNN) and region adaptive synthetic minority oversampling technique algorithm (RA-SMOTE) in view of the low detection rate and high false alarm rate of existing intrusion detection models, and ineffective response to low-frequency attacks. The model achieved a significant increase in the detection rate of low-frequency attacks while improving the overall detection efficiency, and had a certain detection rate for unknown new attacks.

Although machine learning based intrusion detection methods have made good progress, the existing methods for organized attacks [7, 8], such as DDoS attacks, cannot provide comprehensive robust detection [9, 10]. Nazrul Hoque et al. pointed out in [9] that existing DDoS detection methods usually cannot detect low-rate or high-rate DDoS attacks both, and can only detect some types of DDoS attacks, rather than most or even all types of DDoS attacks. In addition, in [10], the authors believed that an ideal defense mechanism requires more nodes in the network to participate in the defense, detection and response of DDoS attacks to deal with the problem that single deployment point detection has poor robustness.

In order to solve the above problems, the graph-based methods [11] have attracted widespread attention from scholars. The network topology is constructed through the communication relationship of IP addresses, and by learning and mining the latent features [12] of the network topology, more accurate detection of organized attacks can be achieved. Network representation learning, also known as network embedding, is a technique that represents the nodes in the network as low-dimensional dense vectors with a certain inference ability [13]. Through network representation learning, the latent features in the network topology are embodied in the form of low-dimensional vectors, so that existing machine learning algorithms such as clustering and neural networks can be used for further detection.

Intrusion behavior is different from normal behavior, so researchers usually use outlier detection algorithms to detect the vectors representing the intrusion behavior in the form of outliers [14]. In [15], the authors proposed an algorithm combining LOF (Local Outlier Factor) and DBSCAN algorithm to detect abnormal network traffic. This algorithm improved the accuracy of network traffic scenarios and reduced the time overhead in the case of large-scale traffic data. Yin Na et al. proposed a network anomaly detection method based on hybrid clustering algorithm (NADHC) in [16],

which combined distance-based clustering algorithm with density-based clustering algorithm to achieve higher detection rate and accuracy, and low false alarm rate.

In this paper, we combine the advantages of network embedding and outlier detection, obtain node vectors of network topology through network embedding technique, and utilize outlier detection algorithm to calculate network embeddings to detect outliers, which are considered to be organized intrusions. In summary, the main contributions and innovations are as follows:

- (1) We propose EmbedLOF, an intrusion detection method for organized attacks based on network embedding. EmbedLOF can perform more comprehensive and robust detection of organized attacks in the network, and malicious nodes corresponding to most types of attacks can be detected.
- (2) We use the Embed algorithm [17] to calculate network embedding. This algorithm greatly reduces the time complexity and space complexity required to obtain network embeddings through  $K + \beta$  Reduction technique, so that the EmbedLOF method can be applied to large-scale networks.
- (3) The EmbedLOF method proposed in this paper has achieved high accuracy and recall when experimented on the ISCX Botnet 2014 dataset [18], which shows the effectiveness of the method.

## 2 Background

This section introduces the background knowledge involved in this article, including some concepts in graph theory, network embedding and LOF outlier detection algorithm.

### 2.1 Graph Theory

This paper involves the concepts of graph, undirected graph, connected graph, and connected component in graph theory. The definitions of these concepts are introduced below [19].

**Definition 1** A **graph**  $G$  is a triplet consisting of a vertex set  $V(G)$ , an edge set  $E(G)$ , and a relationship. The relationship makes each edge related to two vertices (not necessarily different vertices), and these two vertices are called the endpoints of this edge.

**Definition 2** An **undirected graph**  $G$  is a triple, which contains a vertex set  $V(G)$ , an edge set  $E(G)$ , and a function that assigns an unordered pair of vertices to each edge. The edges of an undirected graph have no direction.

**Definition 3** If there is a sequence  $u, e_1, v_1, \dots, v_{k-1}, e_k, v, k \geq 1$  for any  $u, v \in V(G)$ , so that for  $1 \leq i \leq k$ , the endpoints of the edge  $e_i$  are  $v_{i-1}$  and  $v_i$  ( $v_0 = u$ ,  $v_k = v$ ), then the graph  $G$  is said to be **connected**.

**Definition 4** The maximal connected subgraph of graph  $G$  is a connected subgraph of  $G$  which is not included in any other connected subgraphs of  $G$ . The maximal connected subgraph of graph  $G$  is called the **connected component** of graph  $G$ .

## 2.2 Network Embedding

In the traditional network representation, the nodes in the network are related to each other to some extent through the edge set  $E$ . These relationships make calculations in most network processing or analysis algorithms to be iterative or combinatorial, resulting in higher computational complexity [20]. Network embedding aims to learn low-dimensional vector representations of network nodes. Another similar concept is graph embedding, which is also a low-dimensional vector representation of learning nodes. In [20], the author pointed out that the main differences between the two are: (1) the goal of network embedding is to reconstruct the original network and support network inference, while the main goal of graph embedding is to reconstruct the graph; (2) graph embedding is mainly suitable for graphs constructed from feature represented datasets. The proximity between nodes can be well defined by the weights of edges in the original feature space. While network embedding is mainly aimed at naturally formed networks. In this kind of network, the proximity between nodes is not clearly or directly defined. Currently representative network embedding algorithms include DeepWalk [21], Node2vec [22], LINE [23], SDNE [24], etc.

Currently, there are few applications of network embedding in the field of intrusion detection, but there have been attempts in security-related fields to use graph embedding to achieve better performance. In the field of malware detection, the authors in [25] used the function call graph to generate graph embedding, which is used as the input of the deep learning classification model together with API vector to realize the detection of malware. In the field of malicious domain detection, Kai Lei et al. [26] used graph embedding technique to automatically learn the dynamic discriminative feature representation of labeled domains. Based on this, support vector machines were used to achieve malicious domain detection. In [12], the authors obtained the first-order and second-order graph embeddings by optimizing the objective functions of the first-order graph and the second-order graph respectively, and then trained the classifier using these graph embeddings. The experimental results on the CIDDS-001 dataset and the CICIDS 2017 dataset showed that this method could learn latent features better and improved the accuracy of anomaly detection.

In this paper, we use the Embed [17] algorithm to calculate the network embedding, and utilizes the LOF algorithm to further process the network embedding, so as to mine the latent features better and improve the recall of detection.

### 2.3 LOF Outlier Detection Algorithm

The LOF algorithm is a density-based outlier detection algorithm, which is widely used in outlier detection, and there are a large number of improved algorithms based on LOF and variants of LOF [27, 28]. For dataset  $S$ ,  $p, o, o'$  are objects in  $S$ . We use  $d(p, o)$  to denote the distance between  $p$  and  $o$ . Then the LOF algorithm is defined as follows [29]:

**Definition 5** The  $k$ -th distance of object  $p$  is

$$k - \text{distance}(p) = d(p, o), \quad (1)$$

which satisfies:

- (a) There are at least  $k$  objects  $o'$  excluding  $p$  in the set  $S$ , satisfying  $d(p, o') \leq d(p, o)$ ;
- (b) There are at most  $k - 1$  objects  $o'$  excluding  $p$  in the set  $S$ , satisfying  $d(p, o') < d(p, o)$ .

**Definition 6** The  $k$ -th neighborhood of object  $p$  is all objects within the  $k$ -th distance of  $p$ , including the  $k$ -th distance, denoted as  $N_k(p)$ .

**Definition 7** The  $k$ -th reachability distance of object  $p$  with respect to object  $o$ :

$$\text{reach-dist}_k(p, o) = \max\{k - \text{distance}(o), d(p, o)\}. \quad (2)$$

**Definition 8** The local reachability density of object  $p$ :

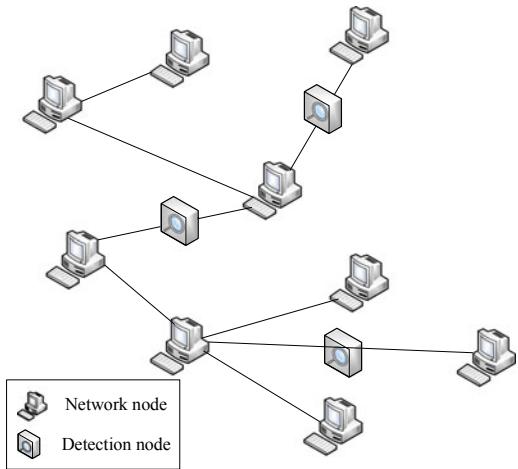
$$\text{lrd}_k(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} \text{reach-dist}_k(p, o)}. \quad (3)$$

**Definition 9** LOF value of object  $p$ :

$$\text{LOF}_k(p) = \frac{\sum_{o \in N_k(p)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(p)}}{|N_k(p)|}. \quad (4)$$

According to the LOF value of each object, given a threshold  $\sigma$ , the objects whose LOF values are greater than the threshold are detected as outliers.

**Fig. 1** Deployment of detection nodes



### 3 Method

This section will introduce the EmbedLOF intrusion detection method in detail, including detection model, preprocess, Embed algorithm and intrusion detection process.

#### 3.1 Detection Model

As shown in Fig. 1, there are a total of  $n$  network nodes and  $w$  detection nodes. The detection nodes capture the pcap data packets communicated between network nodes in the network and construct an undirected network topology graph  $G(V, E)$  about the network nodes, where  $V$  is the set of  $n$  network nodes, and  $E$  is the set of edges between these  $n$  network nodes. In this paper, we will use the constructed network topology to detect the intrusion of organized attacks. Since there are multiple detection nodes in the network, when an organized attack is encountered, a single point of failure can be avoided, thereby enhancing the robustness of intrusion detection.

#### 3.2 Preprocess

The preprocess step first removes the Ethernet frames which do not contain the network layer in the pcap file, so that each packet contains the source IP address and the destination IP address. Then number each IP address starting from 0 as an index. For the source IP address and destination IP address in each packet, we record one and only edge, that is, for source IP address  $sip_j$  and destination IP address  $dip_j$  in

packet  $j$ , the corresponding index of  $sip_j$  is  $a_j$ , the corresponding index of  $dip_j$  is  $b_j$ , if the edges  $(a_j, b_j)$  and  $(b_j, a_j)$  are not in the current edge set  $E$ , then the edge  $(a_j, b_j)$  will be added to  $E$ . After this, the undirected network topology  $G(V, E)$  can be obtained.

Since the Embed algorithm is only applicable to connected graphs, in the preprocessing step, it is also necessary to extract each connected component of the network undirected graph. For each connected component, its node number is renumbered from 0 again.

In the following algorithm, each node is referred to by the number corresponding to its IP address.

### 3.3 *Embed Algorithm*

#### **Algorithm 1: Embed algorithm**

**Input:** Edge set  $E$  of undirected connected graph  $G$ , objective function  $O(\cdot)$ , iteration step size  $\gamma$ , balance factor  $\alpha$ , iteration threshold  $\delta$ , threshold of iteration rounds  $t$

**Output:** Network embedding  $\bar{X}_i = (x_i^1, x_i^2, \dots, x_i^d)$  of node  $i$ , where  $d$  is the number of communities in network

- (a) Initialization: use graph partition software METIS to generate initial partition of graph  $G$ , and initialize the network embedding of node  $i$  as  $Current_i = (x_i^1, x_i^2, \dots, x_i^d)$ , where:

$$x_i^j = \begin{cases} 1/\sqrt{2} & \text{if node } i \text{ belongs to cluster } j \\ 0 & \text{others} \end{cases}$$

- (b) Iteration:
  - (1) For non-edge set  $E_n$ , sample a subset  $E_s$  where  $|E_s| = \alpha |E_n|$
  - (2) Calculate the direction vector  $Direction_i$
  - (3) Calculate new network embedding  $Next_i$  of node  $i$  using  $Next_i = Current_i - \gamma Direction_i$ ;
  - (c) if  $|O(Next_i) - O(Current_i)| / O(Current_i) < \delta$  or iteration rounds exceed  $t$ , end iteration and output  $Current_i$ . Otherwise, let  $Current_i = Next_i$  and repeat (b)

The Embed algorithm is the network embedding algorithm proposed in [17]. The network embedding obtained by this algorithm has the following characteristics:

- (1) The distance between network embeddings of nodes in the same community is similar, that is, about 0;
- (2) The distance between network embeddings of nodes in different communities is close to 1, which is embodied in the large difference in some dimensions of the network embeddings;

- (3) When a node connects to multiple different communities at the same time, there are multiple dimensions which are obviously non-zero in its network embedding. The LOF algorithm is based on this feature to detect nodes that are connected to multiple communities at the same time. Such nodes are usually malicious nodes that control multiple other nodes, such as command and control servers or bot hosts in a botnet.

The Embed algorithm is shown in Algorithm 1. Where the function  $O(\cdot)$  is the objective function in [17], namely

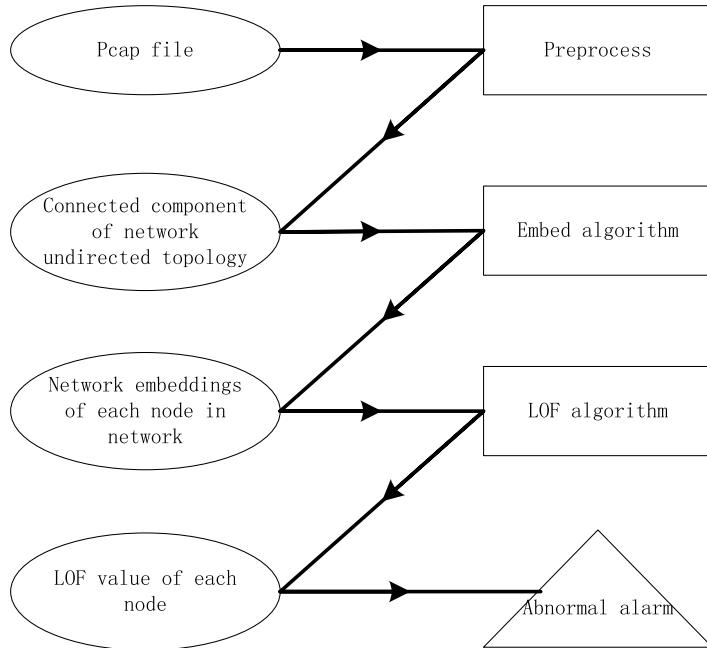
$$O(\overline{X_1}, \overline{X_2}, \dots, \overline{X_n}) = \sum_{(i,j) \in E} \|\overline{X_i} - \overline{X_j}\|^2 + \sum_{(i,j) \in E_s} (\|\overline{X_i} - \overline{X_j}\| - 1)^2. \quad (5)$$

The Embed algorithm proposes the  $K + \beta$  Reduction technique to reduce the computational complexity, where  $K$  is the maximum number of communities connected by abnormal nodes, and  $\beta$  is the tolerance parameter. In each iteration, the network embeddings  $Current_i$  and  $Next_i$  retains the largest  $K + \beta$  dimensions instead of  $d$  dimensions, and the remaining dimensions are set to 0. This greatly reduces the time complexity and space complexity of the algorithm.

### 3.4 Intrusion Detection Process

The algorithm's flow chart is shown in Fig. 2. As shown in Fig. 2, the process of using the EmbedLOF algorithm to detect intrusions is as follows:

- (1) the detection nodes capture the traffic in the network, preprocess the pcap file, and obtain several connected components of the network undirected graph;
- (2) for each connected component, given objective function  $O(\cdot)$ , iteration step size  $\gamma$ , balance factor  $\alpha$ , iteration threshold  $\delta$ , threshold of iteration rounds  $t$ , maximum number of communities connected by abnormal nodes  $K$ , tolerance parameter  $\beta$ , use Embed algorithm to get the corresponding network embedding of each node;
- (3) given  $k$ , calculate the LOF value of each node according to the network embeddings;
- (4) compare the LOF value of each node with a given threshold  $\sigma$ , and mark the node whose LOF value is larger than  $\sigma$  as abnormal and issue an alarm.



**Fig. 2** Flow chart of EmbedLOF algorithm

## 4 Experiments Evaluation

### 4.1 Dataset

In order to verify the detection effect of the EmbedLOF algorithm on organized attacks, we use the ISCX Botnet 2014 dataset from the University of New Brunswick to conduct experiments.

This dataset uses the overlay method to synthesize the ISOT dataset, ISCX 2012 IDS dataset and the botnet traffic generated from the Malware capture facility project, and uses the packet generator BitTwist to map the botnet IPs to hosts outside the current network. Then it uses TCPReplay to replay the malicious and benign traffic and captures all the traffic as a single dataset using TCPdump.

This dataset explicitly lists malicious IP addresses, as shown in Table 1. We conduct experiments on `ISCX_Botnet-Training.pcap` and `ISCX_Botnet-Testing.pcap` in this dataset. For the convenience of explanation, they are denoted as Exp 1 and Exp 2 respectively. After preprocessing these two files, the largest connected components contain 51,951 and 27,654 nodes respectively, and the second largest connected components have no more than 82 nodes. Therefore, only these two largest connected components are used for experiments to illustrate the performance of the EmbedLOF algorithm. Among the 51,951 nodes in Exp 1, 12 nodes

**Table 1** Malicious IP addresses

Attack type	IP address
IRCbot and black hole	10.0.2.15
Neris	147.32.84.180
TBot	172.16.253.130, 172.16.253.131, 172.16.253.129, 172.16.253.240
RBot	147.32.84.170
Menti	147.32.84.150
Sogou	147.32.84.140
Murlo	147.32.84.130
Zero access	172.16.253.132, 192.168.248.165
Virut	147.32.84.160
Zeus	172.29.0.116, 192.168.3.35, 192.168.3.25, 192.168.3.65
Smoke bot	10.37.130.4
Weasel	Botmaster: 74.78.117.238, Bot: 158.65.110.24
Osx_trojan	172.29.0.109
IRC attack	192.168.5.122, 192.168.2.113, 192.168.2.112, 192.168.2.110, 192.168.4.120, 192.168.1.103, 198.164.30.2, 192.168.2.109, 192.168.4.118

correspond to malicious IP addresses; among the 27,654 nodes in Exp 2, 30 nodes correspond to malicious IP addresses.

## 4.2 Evaluation Metrics

The evaluation metrics used in our experiments are accuracy, precision and recall. These evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (8)$$

where  $TP$  represents the number of malicious IP addresses detected,  $TN$  represents the number of normal IP addresses that have not been identified as malicious IP addresses,  $FN$  represents the number of malicious IP addresses that have not been detected, and  $FP$  represents the number of normal IP addresses that have been identified as malicious IP addresses.

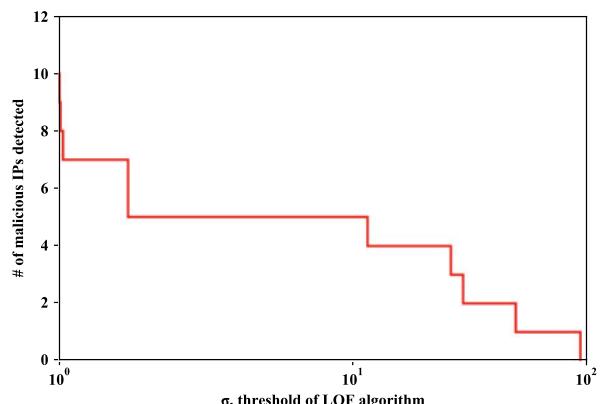
### 4.3 Experiment Results

The Embed algorithm in this paper is based on the C++ language, and the LOF algorithm uses the LocalOutlierFactor function in the python sklearn library [30]. To show the performance of EmbedLOF algorithm specifically, we experiment on the effect the threshold  $\sigma$  of the LOF algorithm, the parameter  $k$ , and the abnormal percentage parameter contamination in the LocalOutlierFactor function. The Embed algorithm uses the default parameters in [17], that is,  $\gamma$  is initially 1, and iteratively updates,  $\alpha|E_n| = |E|$ ,  $\delta$  is 0.001,  $t$  is 50,  $d$  is  $n/500$ ,  $K$  is the average degree of all nodes, i.e.,  $2|E|/n$ , and  $\beta$  is  $\lfloor K/4 \rfloor$ .

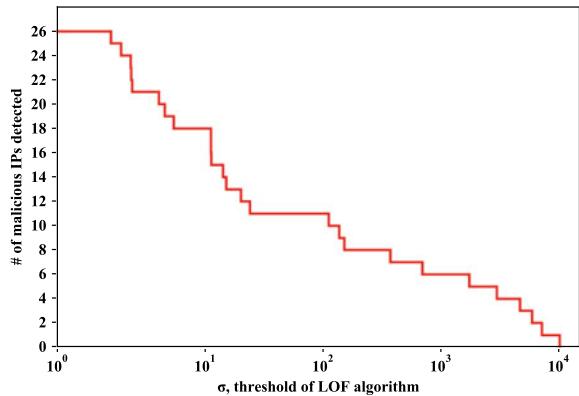
**The Effect of the Threshold  $\sigma$  of the LOF Algorithm.** The LOF algorithm considers the nodes whose LOF values are greater than the threshold  $\sigma$  as outliers. According to the definition of the LOF algorithm, when the LOF value is less than 1, it indicates that the local density of the node is higher than the local density of the node in the neighborhood, and this kind of node is not an outlier; when the LOF value is equal to 1, the local density of the node is close to the local density of the nodes in the neighborhood, and this kind of node is not an outlier. Therefore, this experiment considers the case of  $\sigma \geq 1$ . In Exp 1 and Exp 2,  $k = 5$ , and the contamination parameter is “auto”. The results are shown in Figs. 3 and 4. In order to better display the experimental results, the x-axis in the figures uses the symlog axis, that is, it is linear near 0, and logarithmic in other places. It can be seen from Figs. 3 and 4 that as  $\sigma$  increases, the number of malicious IP addresses detected by the EmbedLOF algorithm decreases. This means that as the requirements for the degree of outlier increase, the malicious behavior of some nodes will be ignored and will not be detected, which is consistent with the general understanding.

**The Effect of the Parameter  $k$  of the LOF Algorithm.** In order to find the appropriate  $k$  to make the algorithm’s detection performance the best, this experiment will take  $k$  from 2 to 29, and observe the number of malicious IP detected, accuracy, precision, and recall of the algorithm. In the experiment,  $\sigma$  is 1.5, and the contamination parameter is “auto”.

**Fig. 3** Effect of threshold  $\sigma$  on number of detected malicious IPs in Exp 1

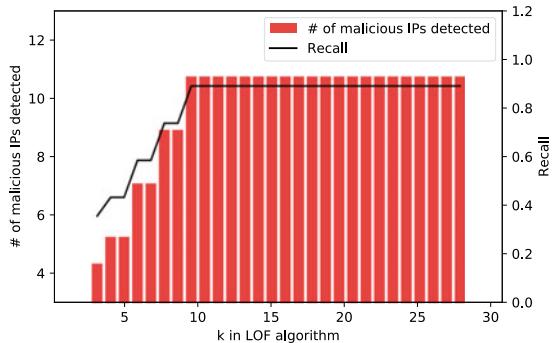


**Fig. 4** Effect of threshold  $\sigma$  on number of detected malicious IPs in Exp 2

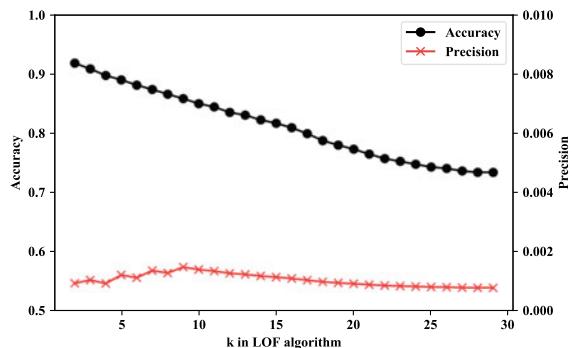


The experimental results are shown in Fig. 5, 6, 7 and 8. It can be seen from Fig. 5 that in Exp 1, when  $k$  is 9 to 29, the EmbedLOF algorithm can detect 91.7% of malicious IPs. It can be seen from Fig. 7 that in Exp 2, when  $k$  is from 11 to 21, the EmbedLOF algorithm can detect 93.3% of malicious IP, and the recall even reaches 96.7% when  $k = 11, 17, 18$ . The high recall means that the various attacks

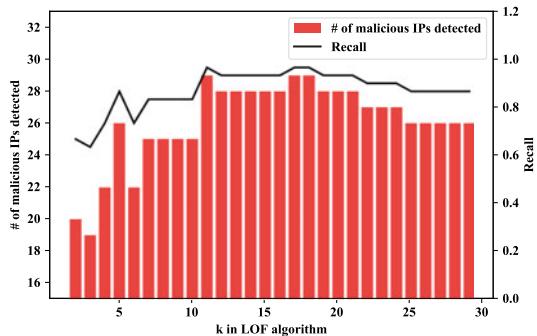
**Fig. 5** Effect of parameter  $k$  on number of detected malicious IPs and recall in Exp 1



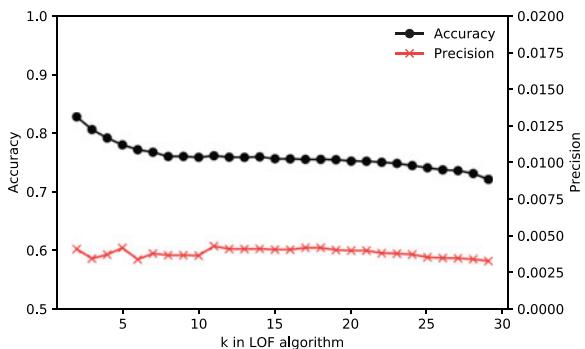
**Fig. 6** Effect of parameter  $k$  on accuracy and precision in Exp 1



**Fig. 7** Effect of parameter k on number of detected malicious IPs and recall in Exp 2



**Fig. 8** Effect of parameter k on accuracy and precision in Exp 2



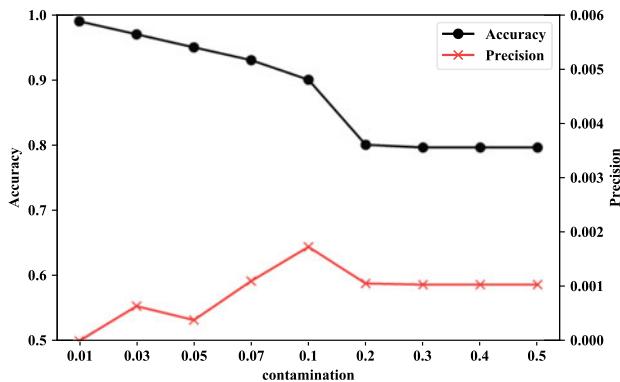
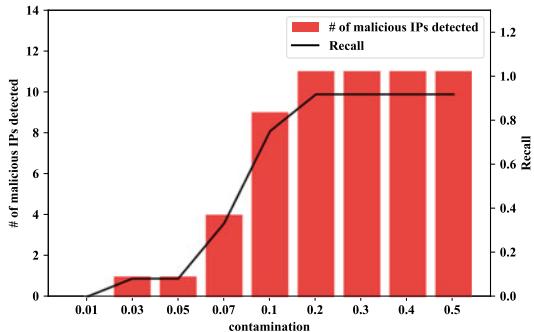
listed in Table 1 can be detected, thus achieving a more comprehensive detection. However, it can be seen from Figs. 6 and 8 that the precisions are very unsatisfactory, which means that the false alarm rate of the algorithm is high, and the real intrusion needs to be found from the alarm. Figures 6 and 8 also show that the accuracy of the algorithm gradually decreases with the increase of k, which means that too much consideration of neighboring nodes interferes with the judgment of the algorithm, so there is a need to make a trade-off between recall and accuracy.

**The Effect of Abnormal Percentage Parameter contamination.** According to the above results, in Exp 1,  $k = 9$  is selected, at this time the recall is the largest, and the accuracy is acceptable. Similarly, in Exp 2, take  $k = 17$ . Take  $\sigma = 1.5$ , and take 0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5 for contamination. The experimental results are shown in Figs. 9, 10, 11 and 12.

It can be seen from Figs. 9 and 11 that the value of contamination has a significant impact on the performance of the EmbedLOF algorithm. If the contamination is too small, most malicious IPs cannot be detected.

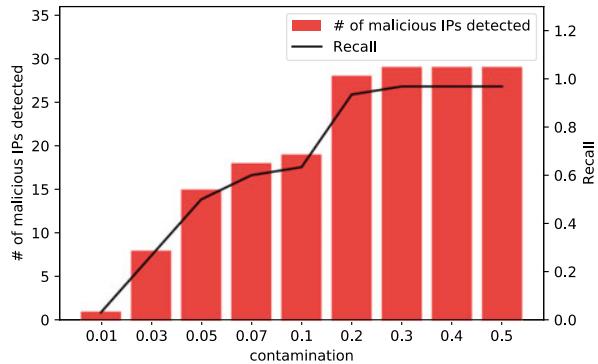
However, as contamination increases, the recall of the algorithm increases, while the accuracy of the algorithm decreases. This can be seen in Figs. 10 and 12. Therefore, when choosing an examination, you need to make a trade-off between accuracy and recall. In addition, it can be seen from Figs. 10 and 12 that the precision of this

**Fig. 9** Effect of parameter ‘contamination’ on number of detected malicious IPs and recall in Exp 1



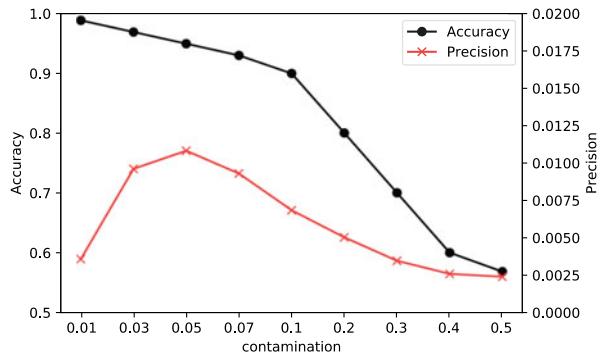
**Fig. 10** Effect of parameter ‘contamination’ on accuracy and precision in Exp 1

**Fig. 11** Effect of parameter ‘contamination’ on number of detected malicious IPs and recall in Exp 2



algorithm is still less than 2% with the change of contamination. This is a shortcoming of this algorithm, and further research is needed.

**Fig. 12** Effect of parameter ‘contamination’ on accuracy and precision in Exp 2



## 5 Conclusion

By combining network embedding with outlier detection, our work makes an attempt to apply network embedding in the field of intrusion detection. The EmbedLOF algorithm proposed in this paper generates low-dimensional vectors from the network topology. At the same time, the latent information about the intrusion in the network contained in the network embedding is expressed in the form of LOF value through the outlier detection algorithm, thus realizing intrusion detection. Experiments on the ISCX Botnet 2014 dataset show that the EmbedLOF algorithm has high recall and accuracy for organized attacks detection. As for future work, we can consider combining network embedding with neural network for intrusion detection.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (Grant No. 62002384), the China Postdoctoral Science Foundation (Grant No. 47698) and the National Natural Science Foundation of China (No. 62072467).

## References

1. CNCERT/CC: National Internet cybersecurity monitoring data analysis report of the first half of 2020. CNCERT/CC, Beijing (2020). [In Chinese]
2. Gümuşbaş, D., et al.: A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. *IEEE Syst. J.* (2020)
3. Drewek-Ossowicka, A., et al.: A survey of neural networks usage for intrusion detection systems. *J. Ambient. Intell. Humaniz. Comput.* **12**, 497–514 (2021)
4. Khraisat, A., et al.: Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity* **2**(20) (2019)
5. Shone, N., et al.: A Deep learning approach to network intrusion detection. *IEEE Trans. Emerg. Top. Comput. Intell.* **2**(1), 41–50 (2018)
6. Yan, B., et al.: Combinatorial intrusion detection model based on deep recurrent neural network and improved SMOTE algorithm. *Chin. J. Netw. Inf. Secur.* **4**(7), 48–59 (2018). In Chinese
7. Bodmer, S., Kilger, M., Carpenter, G., et al.: Reverse Deception: Organized Cyber Threat Counter-Exploitation. McGraw-Hill Education, New York (2012)

8. Huang, K., et al.: Systematically understanding the cyber attack business: a survey. *ACM Comput. Surv.* **51**(4) (2018)
9. Hoque, N., et al.: Botnet in DDoS attacks: trends and challenges. *IEEE Commun. Surv. Tutor.* **17**(4), 2242–2270 (2015)
10. Zargar, S.T., et al.: A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE Commun. Surv. Tutor.* **15**(4), 2046–2069 (2013)
11. Akoglu, L., et al.: graph-based anomaly detection and description: a survey. *Data Min. Knowl. Disc.* **29**(3), 626–688 (2015)
12. Xiao, Q., et al.: Towards network anomaly detection using graph embedding. In: International Conference on Computational Science—ICCS 2020, pp. 156–169. Springer, Cham (2020)
13. Zhang, D., et al.: Network representation learning: a survey. *IEEE Trans. Big Data* **6**(1), 3–28 (2020)
14. Yang, X.: Network Traffic Prediction and Abnormal Traffic Detection Based on Kafka Monitoring System. Beijing University of Posts and Telecommunications (2019). [In Chinese]
15. Gan, Z., et al.: Abnormal network traffic detection based on improved LOF algorithm. In: 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (2018)
16. Yin, N., et al.: Research on application of outlier mining based on hybrid clustering algorithm in anomaly detection. *Comput. Sci.* **44**(5), 122–125, 146 (2017). [In Chinese]
17. Hu, R., et al.: An embedding approach to anomaly detection. In: IEEE International Conference on Data Engineering, pp. 385–396. IEEE, Helsinki, Finland (2016)
18. Beigi, EB., et al.: Towards effective feature selection in machine learning-based botnet detection approaches. In: IEEE Conference on Communications and Network Security, pp. 247–255. IEEE, San Francisco, CA, USA (2014)
19. West, D.B.: Introduction to Graph Theory, 2nd edn. pp. 1–63, Pearson Education (2001)
20. Cui, P., et al.: A survey on network embedding. *IEEE Trans. Knowl. Data Eng.* **31**(5), 833–852 (2019)
21. Perozzi, B., et al.: DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ‘14), pp. 701–710. ACM, New York USA (2014)
22. Grover, A., et al.: node2vec: scalable feature learning for networks. In: the 22nd ACM SIGKDD International Conference, pp. 855–864. ACM, San Francisco, California, USA (2016)
23. Tang, J., et al.: LINE: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1067–1077. International World Wide Web Conferences Steering Committee, Florence Italy (2015)
24. Wang, D., et al.: Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ‘16), pp. 1225–1234. ACM, New York USA (2016)
25. Jiang, H., et al.: DLGraph: Malware Detection Using Deep Learning and Graph Embedding. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1029–1033. IEEE, Orlando, FL, USA (2018)
26. Lei, K., et al.: Detecting malicious domains with behavioral modeling and graph embedding. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), pp. 601–611. IEEE, Dallas, TX, USA (2019)
27. Tang, J., et al.: Enhancing effectiveness of outlier detections for low density patterns. In: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD ‘02), pp. 535–548. Springer, Berlin, Heidelberg (2002)
28. Goldstein, M.: FastLOF: An Expectation-Maximization based Local Outlier Detection Algorithm. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 2282–2285. IEEE, Tsukuba, Japan (2012)
29. Breunig, M.M., et al.: LOF: identifying Density-Based Local Outliers. *ACM SIGMOD Rec.* **29**(2), 93–104 (2000)
30. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)

# An IoT Data Transmission Model Based on Push Mechanism



Deguo Yang · Zeming Wang · Shuai Qi · and Lianqiang Niu

**Abstract** A server is needed to collect the data from the sensors and smart devices. Generally, the end sensors will push data to the server. The traditional server is directly connected with the terminal equipment, and there will be the problem of server bandwidth and connection bottlenecks. To maintain connection and obtain the data from the sensors and smart devices, the mesh, star or tree shape typologies is necessary, and there should be a transmission model pushing data to the server. In this paper, an IoT data transmission model is proposed, and a multi-hop push typology is suggested, and the algorithms are put forward for central server collecting data from the end sensors and devices. The experiment shows that the transmission algorithm is effective for IoT.

**Keywords** IoT · Transmission model · Computer network typology

## 1 Background

IoT (The Internet of Things) is used in many situations. The things include sensors in various environment [1]. The connected architecture is that servers gather many data from sensors. This capability of IoT is used in big data analysis, such as health care, transportation, environment monitoring, farming, etc. [2].

IoT devices use the IEEE 802.15.4, Lora, or NB-IoT standard. The smart devices have the characteristics of less memory capacity, low bandwidth, limited processing capability, short range, etc. [3]. Some nodes act as hosts, and other nodes may as a router. This distinction is that different devices have different capabilities of the device and power available. The standard of 6LoWPAN [4] does not define how topology for IoT networks to be formed. MQTT is a main protocol for IoT collecting data from sensors directly by star topology. The Ad hoc IoT network takes the form of mesh topology. Other application is based on TCP connection, and tree topology

---

D. Yang · Z. Wang · S. Qi · L. Niu

School of Software, Shenyang University of Technology, Shenyang 110023, China

e-mail: [yangdg@sut.edu.cn](mailto:yangdg@sut.edu.cn)

is the best choice. Most of the methods are facing the problem that server bottleneck and failure of single point.

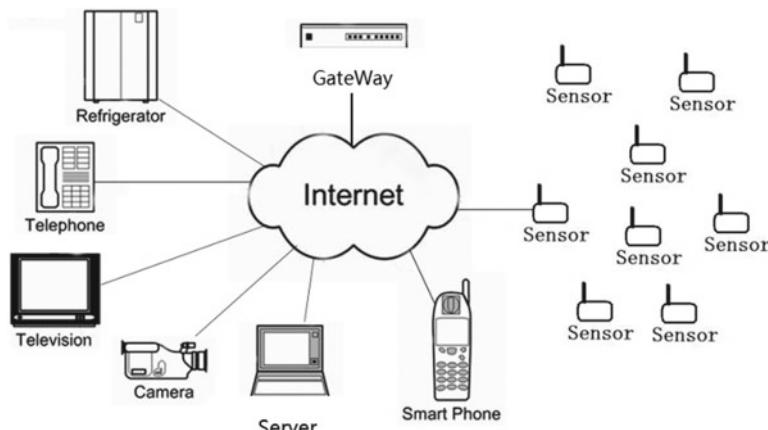
Considering the problems of IoT, some applications of IoT affect the design of the routing requirements. In this paper, a data transmission mechanism based on tree type topology is put forward for IoT applications and a algorithm is proposed to send data to the server [5].

## 2 The Typical Architecture of IoT Topology

Generally, the IoT includes three layers: network layer, perception layer, and transport layer. End sensors should link to the server, and the data is collected and processed by the server.

The IoT applications support the heterogeneous interaction for devices. Two types of network are widely used, i.e. wireless network and wired network. If a main wired network is available, the wired network is the first choice [6]. The wireless devices are connected by the wireless method to the wired network. Because backbone network is used the TCP/IP protocol, it is demonstrated that the interconnection of heterogeneous devices is successful. Generally, devices can be connected in this way, some other sensors can be accessed by wireless method. By the way, some devices can be accessed through the USB port, serial port, and so on. So the gateway is mostly used to connect nonIP networks and IP networks.

A typical IoT architecture example is smart home, which is shown in Fig. 1. It are necessary to connect and control sensors and devices in an environment with a gateway. Gateway can collect all kinds of digital information of smart devices, such as humidity, temperature, light, device switch status. Through the home gateway operating system and a variety of applications can control and monitor all kinds



**Fig. 1** The model of a typical IoT application

of information. Sensors may use the protocol 802.15.4, and others devices can be connected through Ethernet or wireless networks like 802.11g. Some sensors or devices may act as gateways for other devices to provide device interconnection at the network layer. Zigbee, NB-IoT and Lora are new types of low power, are more suitable for the IoT applications.

### 3 The Design of the Model

#### 3.1 *The Model of Measure*

It is defined that the IoT model is based on the assumptions followed. Some devices have enough energy to connect to other intermediary devices, some devices have IP addresses, while some devices do not have IP addresses, such as a large number of sensors and smart devices. Because some devices can only be powered by batteries, such as sensors which are energy-limited. In this model, the residual energy is taken as a value of their ability.

In this paper, intelligent devices can be easily connected to the gateway, such as through USB, serial port or Ethernet. It is convenient for these devices to access the backbone network, which is not described in detail in this paper. The sensors widely used in the Internet of things have the characteristics of limited energy, and can not be directly connected to the network through wired mode. They need to access the network through other intermediate devices, so it becomes very difficult to access the server, especially when the number of sensors is huge. We will emphasize on the large number of sensors network used in the IoT.

Sensor network is different from ad hoc networks, because sensors in sensor networks are usually fixed and not too far away from each other. The data is finally sent to the server, and the position of the sensors for collecting data is relatively fixed. Therefore, we need to construct an effective algorithm and structure for all sensors to build a network to meet the application.

In this design, we use the algorithm similar to the application layer multicast tree to improve the performance of sending data to the server. Tree topology structure is used for sensor data transmission [7].

The primary goal is to form a tree at the application layer. In order to achieve this goal, we need to form a tree structure with server as the root, including all sensors and intermediate gateway. In the process of construction, the data transmission path needs to be optimized. Each end node has paths to the server, and all paths should be chosen to minimize the cost.

We define a node (end sensors and gateways) model and state our assumptions:

- (a) ***MC*** of the node

The capability of the sensors or gateways is important to be selected, so the value should be traded off among many factors, such as capability of process,

energy, delay, and so on. In this paper, the capability from one node to the others is defined as the ***MC***.

In our experiment, full energy of the node is expressed as constant value 1, and the rest of the energy is the percentage of remaining energy. The unit of delay is second. The calculation formula of ***MC*** value is shown in Formula (1).

$$MC = \text{energy} \times p + (1 - \text{delay}) \times (1 - p) \quad (1)$$

The parameter  $p$  is ranged from 0 to 1, and it can be changed in practice.

#### (b) Level of node

If the data of node reach the server through  $n$  intermediate nodes, the node is named  $n$ th level node is marked with ***nodelevel***. The number of nodes in  $n$  ***nodelevel*** marked with  $N_{nl}$ .

### 3.2 Algorithm of Construction Tree

The skeleton of the algorithm is shown in Table 1, and the detail of the algorithm is further elaborated in following section.

The tree is constructed according to the value of every nodes' ***MC*** in this algorithm, and ***MC*** of the nodes can be adjusted by the parameter  $p$  to trade off the relationship of energy and delay.

**Table 1** The construction algorithms for transmission tree

---

It is supposed that there are  $N$  nodes in the application

---



---

$\text{nodenum} = 1; N_{nl} = 1; \text{nodelevel} = 0;$

---



---

0 level node = server;

---



---

Do

---



---

$i = 0, j = 0;$

---



---

Do

---



---

***nodelevel*** node(s) calculate the ***MC*** of its adjacent nodes

---



---

***nodelevel*** node(s) selects the  $L$  ***nodelevel*** + 1 level nodes whose ***MC*** is higher than the rest nodes; ( $L$  is the number of child nodes that the parent node can support)

---



---

$\text{nodenum} = \text{nodenum} + L;$

---



---

$i = i + 1; j = j + L;$

---



---

Until ( $i < N_{nl}$ );

---



---

$N_{nl} = j;$

---



---

$\text{nodelevel} = \text{nodelevel} + 1;$

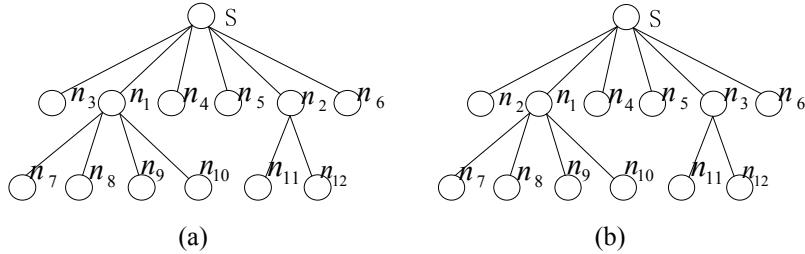
---



---

until ( $\text{nodenum} = N$ );

---



**Fig. 2** The result of algorithm for transmission tree (S stands for server)

We use twelve nodes as an example. The energy of the server is 1, and the energy of  $n_1, n_2$  is 0.9, and the energy of  $n_3, n_4, n_5$  is 0.8, and the energy of the rest nodes are 0.5. We assume that the value of parameter  $p$  in the formula is 0.5. The delay of  $n_1, n_2$  node's from server are 20 ms and 50 ms respectively, and the delay of  $n_3, n_4, n_5$  are 10 ms, 20 ms, 30 ms respectively, and the delay of  $n_6, n_7, n_8, n_9, n_{10}, n_{11}, n_{12}$  are 100 ms, 100 ms, 100 ms, 100 ms, 300 ms, 400 ms, 500 ms respectively.

First, the server calculates the members' **MC**. Taking  $n_1$  as an example, the **MC** of  $n_1$  is calculated as follow.

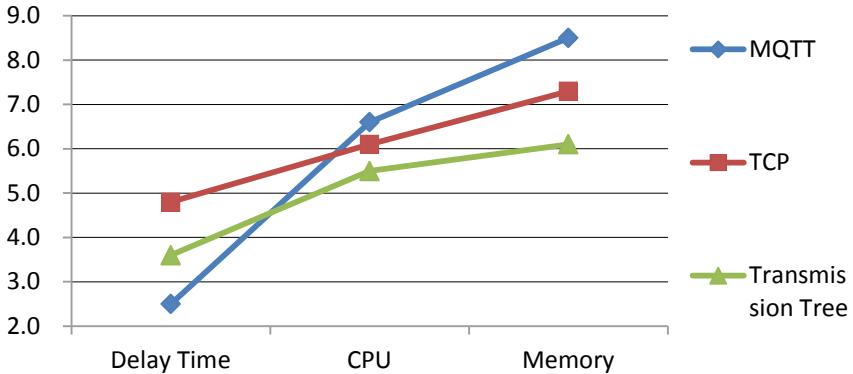
$$MC_{n1} = 0.5 \times 0.9 + (1 - 0.02) \times 0.5 = 0.94. \quad (2)$$

The **MC** of the other nodes  $n_2:0.925, n_3:0.895, n_4:0.89, n_5:0.885, n_6:0.7, n_7:0.7, n_8:0.7, n_9:0.7, n_{10}:0.6, n_{11}:0.55, n_{12}:0.5$ . It is supposed that the server node can support 6 child nodes. The server selects  $n_1:0.94, n_2:0.925, n_3:0.895, n_4:0.89, n_5:0.885, n_6:0.7$  as first level nodes. Then, the nodes  $n_1, n_2$  select the second level nodes. The result of the algorithm is show in Fig. 2a.

If the delay is the more important factor, we can adjust the weight of delay in the formula by reducing  $p$ . For example, the value of  $p$  is set as 0.1, the nodes  $n_1:0.972, n_2:0.954, n_3:0.971$ , and  $n_1, n_3$  will be the two nodes in the second level, and the graph of tree topology is shown in Fig. 2b.

## 4 Experiment Result and Analysis

We used the embedded device made in Bochuang Electronic Technology Co., including a MID (Mobile Internet Device) as the gateway, which is with two USB ports, three serial interfaces. The sensors include one base station board, which is connected with the MID through serial port, ten light sensors, twenty flame sensors, sixty temperature and humidity sensors and six flow sensors. All the sensors are arranged indoor or outdoor. The gateway interacts with the user terminal through Ethernet. The gateway adopts Linux kernel, yaffs2 file system and QT GUI interface.



**Fig. 3** The comparison of experimental results

The user terminal adopts Android operating system. The gateway can collect sensor, camera and voice data to control the intelligent devices in the room. The user terminal can receive information from the gateway server and send information to the gateway for control (Fig. 3).

Three parts of the experiments were carried out. Firstly, we use MQTT protocol to connect all sensors to MQTT server for data collection. Secondly, we use TCP socket method to connect all nodes to the server and transmit data. Finally, we use the method of constructing transmission tree proposed in this paper to connect some nodes to the intermediate nodes, and the intermediate nodes connect to the server through TCP connection. The results show that with the MQTT method and TCP method, the memory consumption (10 k Byte) of server is 1.1 and 1.2 times that of the proposed method, and the CPU consumption (percentage) is 1.4 and 1.2 time of the proposed method in this paper. Three methods are used for data transmission, and the data delay time (ms) is less than 5 ms. The result is shown as It is showed that the method used in this paper can reduce the burden of the central servers and solve the server bottleneck problem to a certain extent, especially when the number of nodes is huge, the effect will be more obvious.

## 5 Summaries

In this paper, an IoT data transmission model is proposed, a server collects the data pushed by sensors and intermediate devices, and a algorithm of transmission tree is proposed. The results of experiment show that the design is feasible and effective.

The future work includes two sides. On one hand, more sensors or intermediate devices should be added to the measurement to construct the transmission tree. On the other hand, it is should be tested that what is the parameter value to achieve the best performance.

## References

1. Harwahyu, R., Cheng, R.G., Sari, R.F.: Performance of distributed coordination function for serving IoT applications. In: 3rd International Conference on Smart City Innovation. FiJi (2020)
2. Pavan, B.S., Mahesh, M., Govindan, H.: Performance anomaly of Group-Synchronized Distributed Coordination Function in IEEE 802.11 ah based Multi-rate IoT Networks. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS) Shanghai (2020)
3. Hui, J., Thubert, P.: Compression Format for IPv6 Datagrams in Low Power and Lossy Networks (6LoWPAN), RFC draft. <http://www.ietf.org/id/draft-ietf-6lowpan-hc-15.txt>.
4. Kushalnagar, N., Montenegro, G.: IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): Overview, Assumptions, Problem Statement, and Goals, RFC draft. <http://datatracker.ietf.org/doc/rfc4919>
5. Yang, D., Zhang, H.: A Design of the Model for Smart Home Gateway. <https://www.atlantis-press.com/proceedings/iccasim-12/2577>
6. Liu, B., Towsley, D.: Capacity of a wireless ad hoc network with infrastructure. In: MobiHoc '07 Proceedings of the 8th ACM International Symposium on Mobile Ad hoc Networking and Computing. ACM, New York ( 2007)
7. Yang, D., Zhao, Y., Wang, C., Gao, Y.: ALMSC: an Application layer multicast based SIP conference model. In: 2006 10th International Conference on Computer Supported Cooperative Work in Design (2006)

# Electrical Load Forecasting Using Hybrid of Extreme Gradient Boosting and Light Gradient Boosting Machine



Eric Nziyumva , Rong Hu , Chih-Yu Hsu , and Jovial Niyogisubizo

**Abstract** Ensemble learning methods have been used to improve performance accuracy through bias-variance trade-off techniques. However, there is still room to improve. This paper proposes an ensemble model to forecast the electrical load behavior based on a hybrid of Extreme Gradient Boosting (XGBoost) and Light gradient boosting machine (LGBM). Extreme gradient boosting (XGBoost), a Light gradient boosting machine (LGBM) and a hybrid of XGBoost and LGBM models are trained, evaluated, and compared. The experiments show that the proposed model outperforms other methods by reducing more than 1% in mean absolute percentage error (MAPE), root mean squared percentage error (RMSPE), and mean absolute error (MAE). The dataset from the Pennsylvania-New Jersey-Maryland interconnection power grid was used to validate the evolutionary capability of the proposed method and the finding of optimal accuracy of the model.

**Keywords** Electrical load forecasting · Ensemble learning · Extreme gradient boosting machine (XGBoost) · Light gradient boosting machine (LGBM)

## 1 Introduction

Energy power prediction is very important in our daily life. It is the first approach to the best power system management and plays a great significance for all-electric power-related activities. Moreover, this prediction not only presents its strength to the reliable grid operation but also for safe electricity planning, modern transportation, communication and has a great positive impact on national security. Thus, accurate electric load prediction is essential for power systems since accurate prediction leads to the economic development of any country through substantial savings

---

E. Nziyumva · J. Niyogisubizo

Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, China

R. Hu

Fujian Provincial Key Laboratory of Big Data Mining, Fujian University of Technology, Fuzhou, China

e-mail: [hurong@fjut.edu.cn](mailto:hurong@fjut.edu.cn)

in operating and maintenance costs [1]. However, achieving the desired accuracy is difficult due to the various factors influencing the electric load behavior include human social activities, country policies, climate change, and economic development [2].

Previously, electric load forecasting (ELF) had been almost entirely limited to traditional statistical methods. The classical researches proposed include the adaptive time-series auto-regressive moving average (ARMA) model presented a good performance of reducing the error compared to the other models. Due to its simplicity and effectiveness, ARMA was popular and extensively used in ELF researches however it is limited to only being used for stationary time-series data. Cheng-Ming Lee and Chia-Nan Ko [3] proposed a new hybrid algorithm based on auto-regressive integrated moving average (ARIMA) which has the advantage of introducing non-stationary time-series data. Compared to other models used in their work, the simulation results indicated also the highest forecasting performance. Unfortunately, the random noise which disturbs the whole process, ARMA and ARIMA models use only time and load as input data which implies the ARMAX and ARIMAX to be discovered for introducing the exogenous variables. Then, Indian researchers Shilpa G N and Dr. G S Sheshadri had used the ARIMAX model, an extension of ARIMA with an exogenous variable for ELF [4]. However, the classical models are limited due to only focus on the relationship between the dependent and independent variables. In addition, their forecasting accuracy is not good enough therefore the modern models were introduced for making the most possible accurate predictions.

With modern science progress, load prediction technologies have been considerably developed. Lately, the introduction of machine learning (ML) theories in the electrical power engineering field became more and more popular which implies great efficiency for improving the performance of forecasting models [2]. The widely used ML models include multiple linear regression (MLR) applied for ELF and gave successful results. Even if it is easy for results interpretation, the regression models may lead to erroneous and misleading results due to the wrong assumptions [5]. Salkuti, S. R. proposed an ANN-based hybrid for predicting short-term electrical load demand in which a better result was found. Besides, to avoid different drawbacks of ANNs such as falling into the trap of local minima during the parameter optimization process, Salkuti, S. R. used a hybrid approach of combining ANN, wavelet transforms (WTs), and evolutionary-based differential evolution algorithm [6]. On the other hand, Mohamed proposed a full wavelet packet transform and neural network-based ensemble method. The simulation results show that the proposed approach reduces MAPE by 20% in comparison with the traditional neural network method [7]. Later, Chengdong developed a wavelet transforms-based model in which the proposed method combines the fuzzy inference system and the periodicity knowledge to generate accurate forecasting results [8]. Due to its double major advantages of being used in optimization and prediction fields, the genetic algorithm (GA) has been well-suited with nonlinear systems and it conducts a particular optimization based on the natural selection of the optimal solutions found from a wide range of forecasting model candidates' population [9]. Then, expert systems-based models had been increasingly developed for handling prediction issues.

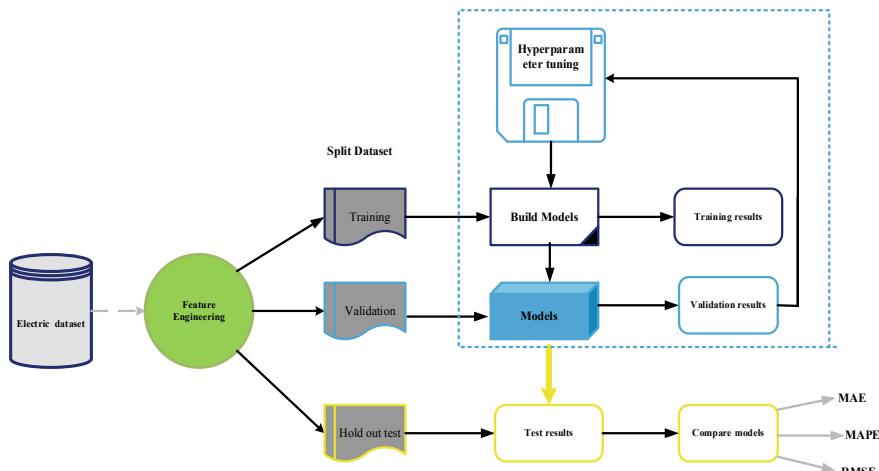
Although the machine learning models had been widely used in various forecasting issues include ELF, their models' performance present various gaps of erroneous results due to variance, bias, and noise. This affects the ELF which leads to significant losses due to maintenance costs, unsafe power system operation, and all power planning-related activities [10]. Moreover, inaccurate load forecasting has great negative impacts on energy generating capacity scheduling which leads to inadequate operating.

In this paper, the proposed approach aims to upgrade the forecasting performance of machine learning algorithms. This is achieved by reducing the error between actual and predicted values through the bias-variance trade-off. The main principle behind this work is the combination of ensemble learning. At first, the extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), Adaptive boosting (AdaBoost), and random forest are firstly compared according to their accuracy and training time. Then, the hybrid of XGBoost-LGBM has been done to enhance the performance accuracy. The innovations of the proposed approach are such as the combination of two models and performance improvement compared to the remaining models used in this paper which leads to significant loss reductions.

The rest of this paper is organized as follows: Sect. 2 describes the methods used in this paper. Section 3 evaluates, discusses, and compares the performance results of models. Section 4 concludes the paper.

## 2 Methodology

The methodology used in this paper is graphically represented through Fig. 1.



**Fig. 1** Overall design of the study

According to their performance accuracy and training time, the boosting-based models include XGBoost, LGBM, Adaboost, and random forest, a bagging-based ensemble learning model, are compared with a hybrid of XGBoost-LGBM, the proposed model. The hyperparameter tuning of the models has been also computed. This section explains the research methodology process. Finally, the performance comparison of the models is done.

## 2.1 Overall Research Design

Figure 1 represents the development of the overall research design proposed to conduct the study for enhancing electric load forecasting. After identifying the inaccurate forecasting problems, the helpful steps of overcoming them have been proposed as follows: First, the electric dataset was gained. Then, feature engineering was conducted. Third, the ensemble machine learning-based techniques (Random Forest, XGBoost, LightGBM, Adaboost, and XGBoost-LGBM) were trained, evaluated, and compared. The mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square percentage error (RMSPE) were analyzed.

## 2.2 Description of Dataset

In this paper, the dataset used comes from Pennsylvania-New Jersey-Maryland (PJM) website [11]. The extracted information holds data of 87,600 samples with a sampling frequency of 1 h. The database covers a range of times starting from July, 1st 2008, and the hourly load demand recorded from twelve electrical networks is expressed in megawatts (MW). Various important techniques have been applied to this dataset for improving the outcome of the models. The first technique is dataset filtering done for selecting the sub dataset to be used for viewing and analysis. The second concerns dataset training which has the role of training the algorithm so that it can predict accurately. The third technique emphasis the evaluation of the dataset which is considered as the performance comparison tasks done with the help of a validation dataset to find the smallest error network. A model may overfit during the validation procedure therefore the performance evaluation might be done to the testing dataset. Cross-validation as the technique of evaluating the models for limited data by resampling had been used. The attributes after data filtering include the day of the week, holidays, season, the hour of the day, month, and energy consumption. The target variable is the consumption of electrical energy expressed in megawatts (MW).

## 2.3 Feature Engineering

Feature engineering expressed as the way of extracting variables from raw data plays an important role in determining the key variables that will be useful to win the upcoming applications and then to classify them as either low and high according to their impacts. Since ML framework development is a process that begins by carefully defining the requirements [12], the iterative process starts which involves building and testing various models over a dataset. To explore and analyze the information, the data gained from the dataset should be pre-processed and transformed. In the end, the relationship between independent features such as the day of the week, holidays, season, the hour of the day, month, year, and target variable (consumption of electrical energy) is seen.

The next stage is model building to try and evaluate different models. Here, the data is organized into three different split sets. With the help of the training and validation sets, there is an optimization possibility of model parameters using cross-validation procedures then hyperparameter tuning. The third set namely the “hold-out test” is used for final testing and model comparison.

## 2.4 Ensemble-Based Machine Learning Models

With the fast improvement of machine learning, various techniques to increase accuracy have been proposed. All the previous works have been done to reduce the errors. Here, we present a brief description of ensemble-based machine learning techniques used in this paper.

**Random forest:** The random forest algorithm has been firstly invented by Tin Kam Ho using attribute bootstrap aggregating (bagging) in 1995 [13]. The functionality of random forest consists of three main parts. Firstly, the samples are selected through the bagging techniques which are used for extracting the N (number) times training datasets from original data. Then, the prediction from each tree-based learner is found. Finally, the result is found through the combination methods such as averaging or voting. Random forests algorithm works very well compared to the decision tree as it corrects the overfitting but its accuracy is lower than that of the gradient boosted trees [14].

**Adaptive Boosting (AdaBoost):** Adaptive boosting is the first boosting-based algorithm developed by the joint of Freund and Schapire [15]. The class of boosting algorithm takes its description as the machine learning approach to increase the forecasting performance level based on the combination of various weak learners and inaccurate rules. As the first practical boosting algorithm, adaptive boosting is widely used and studied and then applied in numerous fields. Its advantages were seen in regression and classification issues handling.

**Extreme Gradient Boosting (XGBoost):** Through the research project conducted by Tianqi Chen, the XGBoost that works under the principle of boosting gradient

tree had officially come out on March 27, 2014 [16]. This model could be used for handling regression or classification issues. Mathematically, the gain leads to a regularized boosting techniques is defined by [17]:

$$Gain = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \beta} + \frac{G_R^2}{H_R + \beta} - \frac{(G_R + G_L)^2}{H_R + H_L + \beta} \right] - \alpha \quad (1)$$

where the first term is the score of the left child, the second the score of the right child and the third the score if we do not split.  $\beta$  and  $\alpha$  are ridge and lasso regularization coefficients respectively.

**Light Gradient Boosting Machine (LGBM):** The Gradient-Based One-Side Sampling (GOSS) decreases the computation costs since it uses a subset of smaller instances rather than using all instances. Exclusive Feature Bundling (EFB) converts exclusive features into less dense features. The combination of both GOSS and EFB techniques produces LGBM [18]. Compared to the traditional gradient boosting concept, the LGBM has an accelerated training process and higher performance accuracy. Furthermore, this kind of gradient boosting can deal with the large dataset since it supports GPU and parallel learning [10].

**Hybrid of XGBoost-LGBM:** The hybrid of XGBoost and LGBM, boosted-based ML models, in which the individual's components are sequentially coupled for building a powerful meta-learner. The idea behind the proposed model is to find the approach of reducing the contribution of both bias and variance to error since the errors and predictions in any ML models are adversely influenced by bias, variance, and noise [19]. A high bias and variance bring about the underfitting and overfitting of the training data respectively while noise is considered as an irreducible error caused by improper cleaning of data. The proposed approach is found through grouping the individual models in sequential. The numerical simulations with cross-validation (CV) and hyperparameters tuning verify the power of the new meta-learner algorithm by giving the best results compared to the single model as shown in Table 1. The expected prediction error of a regression model using squared-error loss is expressed as:

**Table 1** Performance evaluation of models: LGBM, XGBoost and Random forest

CV	LGBM			XGBoost			Random forest		
	MAPE	RMSPE	MAE	MAPE	RMSPE	MAE	MAPE	RMSPE	MAE
K1	1.88	2.12	199.66	1.94	2.04	200.1	2.16	2.58	205.86
K2	1.94	2.24	188.64	1.98	2.56	199.72	2.12	2.84	205.38
K3	1.78	1.94	189.30	1.84	2.52	199.04	2.14	2.92	200.72
K4	1.74	1.96	195.78	1.90	2.48	198.72	1.96	2.70	200.36
K5	1.72	1.78	193.72	1.74	2.46	197.84	1.90	2.66	220.04
K6	1.46	1.68	188.10	1.64	2.24	197.82	1.88	2.54	200.02
Mean	1.74	1.94	192.52	1.84	2.38	198.87	2.02	2.70	205.38

$$\begin{aligned}
 \Psi(y, \hat{F}(x_i)) &= E\left[\left(y - \hat{F}(x_i)\right)^2 | x = x_i\right] \\
 &= \sigma_{\varepsilon}^2 + \left[E\left(\hat{F}(x_i)\right) - F(x_i)\right]^2 - E\left[\hat{F}(x_i) - E\left(\hat{F}(x_i)\right)\right]^2 \\
 &= \sigma_{\varepsilon}^2 + bias^2(\hat{F}(x_i)) + variance(\hat{F}(x_i))
 \end{aligned} \tag{2}$$

where the first term represents an irreducible error, the second term is the contribution of squared bias to error while the last term is the contribution of variance to error.  $F$  and  $\hat{F}$  represent the actual and predicted values respectively while  $E$  represents expected values then  $x = x_i$  represents point value.

## 2.5 Evaluation Criteria

To evaluate and compare the performance of ML models, several performance metrics are used. The mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared percentage error (RMSPE) have been utilized. Mathematically, they are defined:

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - \hat{Y}_i| \tag{3}$$

$$MAPE = \frac{100}{N} * \sum_{i=1}^N \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \tag{4}$$

$$RMSPE = 100 * \sqrt{\frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N Y_i}} \tag{5}$$

$Y_i$  and  $\hat{Y}_i$  are the actual and predicted values respectively while  $N$  is the observation number. The more the values of MAE, MAPE, and RMSPE, the worse prediction accuracy.

## 3 Results and Discussions

The performance evaluation of four ensemble learning-based methods was judged and compared with the proposed model. In this section, the results are presented and discussed.

**Tabel 2** Performance evaluation of models: AdaBoost and XGBoost-LGBM

	AdaBoost			XGBoost-LGBM		
CV	MAPE	RMSPE	MAE	MAPE	RMSPE	MAE
K1	2.14	2.44	202.24	1.78	1.92	199.06
K2	2.10	2.78	203.76	1.70	1.30	196.04
K3	2.06	2.76	200.32	1.62	1.44	188.62
K4	1.92	2.64	200.04	1.50	1.28	184.68
K5	1.86	2.56	199.7	1.46	1.08	182.04
K6	1.82	2.38	199.06	1.24	0.90	178.52
Mean	1.98	2.58	200.84	1.54	1.32	188.16

### 3.1 Comparison

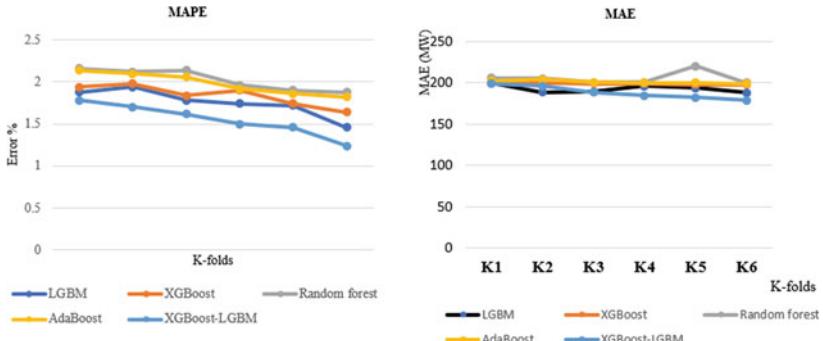
The ML performance metrics are often used to compare the models for selecting the best suitable for ELF. According to the literature, different models have been proposed and compared based on the performance evaluation results with the execution time. Since the existing models have not yet satisfied the desired forecasting quality, the research is still undergoing. Here, the hybrid of XGBoost-LGBM and four single models have been trained, tested, and then compared. For making an accurate performance evaluation, the simulation process is repeated six times. The technique used here is popularly known as k-fold cross-validation (CV) which prevents the model to overfit the new data [20]. Then, hyperparameter tuning has been also applied.

The six-fold cross-validation results for each model are resumed in Tables 1 and 2. The performance results before applying CV techniques are worse than those obtained after using it. The errors were higher which indicates significant losses due to the inaccurate electric load prediction.

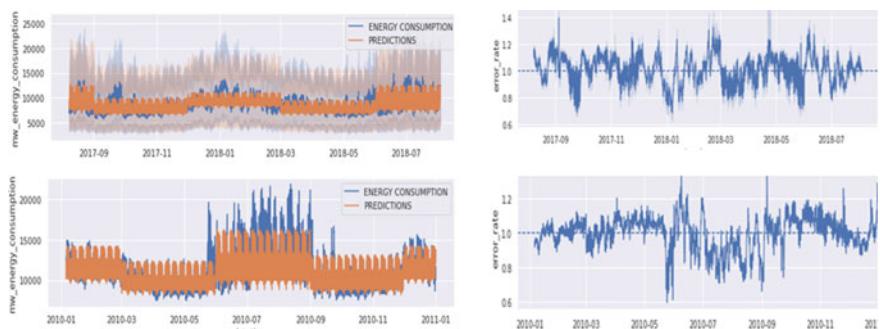
According to the mean absolute percentage error (MAPE) which is considered as a loss function to define the error termed by the model evaluation. The mean results for the k-folds cross-validation reveal that the proposed hybrid successfully limits the forecasting error to 1.54% compared to the other remaining single models.

Figure 2 represents the comparison of the MAPE and MAE obtained from various models used in this paper. The proposed model, a hybrid of XGBoost and LGBM, shows the highest accuracy compared to the other remaining models.

In addition, the root mean squared percentage error (RMSPE) of the single models is greater than that of the proposed model. Based on the experiment results, the hybrid of XGBoost-LGBM reduces the error to 1.32%. Afterwards, the mean absolute error (MAE) expressed as the measure of how far the predictions were from the actual output shows that the proposed model has the least values of error. Based on the aforementioned discussions and performance metrics principle (the smaller the values of MAPE, RMSPE and MAE, the better the model performance) therefore it is reasonable and proves that the proposed model outperforms the other models



**Fig. 2** Comparison of the MAPE and MAE obtained



**Fig. 3** Energy consumption versus Estimation and Error rate versus datetime

tested in this paper. However, the proposed model takes a longer time to be trained compared to single models.

Figure 3 shows ELF with two months ahead. The left graphs represent the energy consumption and prediction versus datetime while the right graphs show the error rate versus datetime. For the first-row graphs, there are fewer losses as the actual energy consumption is almost equal to the predicted electrical energy therefore the utilization of generated energy is maximized. While for the second-row graphs, the significant error values are greatly remarkable which indicates a significant difference between predicted and actual values. Consequently, the first row-graphs show an accurate model which leads to the main goal of ELF.

## 4 Conclusion

The overall development of any country is based on proper load forecasting. The inaccurate forecasting affects negatively the planning and may lead to various significant

losses. Hence, the performance accuracy should be improved. This paper proposes the techniques of enhancing accuracy for electric load forecasting. The advantages of both XGBoost and LGBM are combined for achieving the target. The proposed model, a hybrid of XGBoost and LGBM, shows the highest accuracy compared to the other remaining models tested in this paper. According to the mean absolute percentage error (MAPE), the mean results for the k-folds cross-validation reveal that the proposed hybrid successfully limits the forecasting error to 1.54% compared to the other remaining single models. In addition, the root mean squared percentage error (RMSPE) of the single models are greater than that of the proposed model which is not good. On the other hand, the mean absolute error (MAE) shows that the proposed model has the least values of prediction error which leads to attractive results. Moreover, the accurate forecasting obtained from the proposed approach leads to the reduction of the significant losses since the utilization of power-generated energy is maximized. Furthermore, the contributions of this proposed approach include safe power system operation, proper planning of transmission and distribution facilities, proper financing (future expenditure and earnings), substantial savings in operating and maintenance costs, and then safe power planning related activities while its innovations are such as the combination of two models and performance improvement compared to the other models tested in this experiment. Based on the aforementioned discussions and performance metrics principle (the smaller the values of MAPE, RMSPE, and MAE, the better the model performance) therefore it is reasonable and proves that the proposed model outperforms the other models tested in this paper.

## References

1. Jung, S.-M., Park, S., Jung, S.-W., Hwang, E.: Monthly electric load forecasting using transfer learning for smart cities. *Sustainability* **12**(16), 6364 (2020)
2. Wang, R., Wang, J., Xu, Y.: A novel combined model based on hybrid optimization algorithm for electrical load forecasting. *Appl. Soft Comput.* **82**, 105548 (2019)
3. Lee, C.-M., & Ko, C.-N.: Short-term load forecasting using lifting scheme and ARIMA models. *Expert Syst. Appl.* **5902–5911** (2011)
4. Shilpa, G.N., Sheshadri, G.S.: ARIMAX model for short-term electrical load forecasting. *Int. J. Recent Technol. Eng. (IJRTE)* **8**(4) (2019)
5. Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., Torres, J.: stacking ensemble learning for short-term electricity consumption forecasting. *Energies* **11**(4), 949 (2018)
6. Salkuti, S. R.: Short-term electrical load forecasting using hybrid ANN–DE and wavelet transforms approach. *Electr Eng* **100**(4), 2755–2763 (2018)
7. El-Hendawia, M., Wang, Z.: An ensemble method of full wavelet packet transform and neural network for short term electrical load forecasting. *Electr. Power Syst. Res.* **182**, 106265 (2020)
8. Li, C., Tang, M., Zhang, G., Wang, R., Tian, C.: A hybrid short-term building electrical load forecasting. *Int. J. Fuzzy Syst.* **22**(1), 156–171 (2020)
9. Al-Douri, Y.K., Al-Chalabi, H., Lundberg, J.: Time Series forecasting using genetic algorithm. In: The Twelfth International Conference on Advanced Engineering Computing and Applications in Sciences (2018)
10. Quinto, B.: Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More (2020)

11. PJM.PJM load forecast, [Online]. Available <https://dataminer2.pjm.com/list>. Accessed 11 Nov 2020
12. Al Mamun, A., Sohel, M., Mohammad, N., Sunny, M.S.H., Dipta, D.R., Hossain, E.: A comprehensive review of the load forecasting techniques using single and hybrid predictive models. *IEEE Access* 8, 134911–134939 (2020)
13. Ho, T.K.: Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition* 1, 278–282 (1995)
14. Piryonesi, S. M., El-Diraby, T. E.: Role of data analytics in infrastructure asset management: overcoming data size and quality problems. *J. Transp. Eng. Part B: Pavements* **146**(2), 04020022 (2020)
15. Schapire, R.E.: Explaining AdaBoost. *Empirical Infer.* 37–52 (2013)
16. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco (2016)
17. Omar, K.B.A.: Xgboost and LGBM for Porto Seguros Kaggle Challenge: A Comparison (2018)
18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: a highly efficient gradient boosting decision tree. In: *Proceedings of the Advances in Neural Information Processing Systems* 30, 3146–3154 (2017)
19. Géron, A.: *Hands-on machine learning with Scikit-Learn, Keras*, Canada: O'Reilly Media, Inc., Sebastopol 1492032611 (2019)
20. Lin, Y., Luo, H., Wang, D., Guo, H., Zhu, K.: An ensemble model based on machine learning methods and data preprocessing for short-term electric load forecasting. *Energies* **10**(8), 1186 (2017)

# Research on Splicing of Horizontal and Longitudinal Shredded Paper Based on Hungarian Algorithm



Lizhi Shen , Zhixiong He , Lang Chen , and Rongyuan Chen

**Abstract** In order to solve the problems of traditional heuristic algorithm for solving the splicing of horizontal and longitudinal shredded paper, such as high complexity and low success rate when the order of magnitude is large. A method is proposed to transform the horizontal and longitudinal shredded paper splicing problem into an assignment problem. Then the Hungarian algorithm is used to solve the problem: first, the fragments are numbered and the position information on both sides of the fragments is extracted. For the shredded paper with both crosscutting and longitudinal cutting. The K-Means clustering algorithm is used to splice each row successfully. Then, one side of any fragment is regarded as the person completing the task, and the other side is regarded as the task, and the gray difference of the corresponding position on both sides is regarded as the cost for the person to complete the task (if the blank row is cut, the line spacing difference is used as the cost). Thus the cost matrix is constructed. Finally, through the minimum cost sum obtained by the row-column transformation of the Hungarian algorithm, the optimal solution of the problem is obtained and the fragment splicing sequence is obtained. This method transforms the nature of the problem ideologically and adopts a solution different from the traditional heuristic algorithm. Compared with the simulated annealing algorithm, the complexity of the Hungarian algorithm is obviously lower than that of the simulated annealing algorithm. The amount of calculation is greatly reduced.

**Keywords** Shredding problem · Assignment model · Hungarian algorithm · Simulated annealing algorithm · Clustering algorithm

## 1 Introduction

The restoration of damaged images has important applications in the fields of judicial material evidence restoration, historical literature restoration, banknotes and military

---

L. Shen · Z. He · L. Chen ( ) · R. Chen  
Hunan University of Technology and Business, Changsha, China  
e-mail: [2208@hutb.edu.cn](mailto:2208@hutb.edu.cn)

intelligence acquisition. In the past, the traditional splicing repair work was basically completed manually, the accuracy is very high, but the efficiency is very low. Especially when the scale of the problem is enormous, it is difficult for manpower to complete the task in a short period of time. Therefore, the research of using computers to realize automatic splicing of fragments is of great practical significance [1–3].

At present, scholars at home and abroad mainly classify the splicing and repair of text fragments into two categories: irregular and regular fragments. For the repair of irregular fragments, the frequently used algorithms are boundary detection algorithm, genetic algorithm and so on [4]. Xu Muhan et al. aiming at the problems of poor adaptability and low success rate of automatic splicing of irregular fragments based on machine vision, a new edge shape descriptor method and a multi-dimensional matching irregular fragment stitching method combined with color information are proposed [5]. Dang Yuechen et al. proposed a shredded paper splicing reconstruction algorithm based on evolutionary computation in view of the difficulty of manual splicing [6]. Luo Zhizhong et al. studied the splicing repair of shredded paper from the point of view of text features. They mainly make use of the edge shape outline and edge grayscale information of the fragments. The main methods used to repair fragments with regular shape are 0–1 programming, clustering algorithm, ant colony algorithm, simulated annealing algorithm and so on [7]. Zhuang Sifa and Fu Ximei proposed a horizontal sorting restoration algorithm based on 0–1 programming for regular paper fragments cut horizontally and longitudinally, and a classification algorithm for the case of both crosscutting and vertical cutting [8]. Liu Qiuju et al. proposed a fragment splicing restoration method based on text features by establishing a dynamic programming model [9]. Fu Guanghui et al. in view of the fact that the existing algorithms are not robust enough and the accuracy of intra-line splicing of fragments is low, an improved method of automatic splicing of paper fragments based on clustering and ant colony algorithm is proposed [10]. Han Yingying et al. and Lan Yang et al. realize splicing by establishing a splicing model based on 0–1 programming [11, 12]. Zhong uses Q clustering analysis to splice the regular fragments by constructing edge contrast matrix [13]; Zhou transforms the problem into traveling Salesman problem and proposes improved genetic algorithm and greedy algorithm to solve the regular fragments [14]; Liang et al. uses degenerate algorithm to solve the regular fragments by constructing the corresponding gray matrix and deducing the edge characteristics [15].

In the above research, many scholars regard the horizontal and vertical shredding paper splicing problem as a traveling salesman problem or integer programming problem, usually utilizing traditional heuristic algorithms such as ant colony algorithm, simulated annealing algorithm and so on. In this paper, the Chinese horizontal and vertical shredding paper splicing problem is transformed into the assignment problem in operational research, and the minimum cost mathematical model is established and solved by the Hungarian algorithm.

## 2 Problem Modeling

### 2.1 Problem Description

In the above two sections, the properties of the assignment problem and how to use the Hungarian algorithm to solve the assignment problem are described in detail. Going back to the existing longitudinal shredding problem, we need a series of transformations to convert it into an assignment problem. After numbering  $n$  pieces of paper, the right side of each piece of paper is defined as the task  $a_k$  to be completed, and the left side of each piece of paper is defined as the person who completed the task  $b_k$ . At the same time, the position information of the left and right sides of each pair of shredders is extracted and expressed as a position vector, the gray difference between the left side of the  $i$  shredder and the right side of the  $j$  shredder means that the cost of completing the  $j$  task is  $c_{1ij}$ . From this we can get the cost matrix  $C1 = (c_{1ij})$ .

For the problem of horizontal and vertical shredding paper, the cost matrix is divided into two parts: grayscale difference and line spacing difference. After the rows are spliced successfully by K-means clustering algorithm, only the upper side of each piece of paper is defined as the task to be completed, and the lower side of each piece of paper is defined as the person who completes the task. At the same time, considering the case of skimming and pressing in the font stroke, the gray difference between each pixel on the lower side of the  $i$  shredder and the upper side of the  $j$  shredder and the gray difference between the three pixels before and after the matching is compared. If the match is successful, the value will be increased by 1, and the value of the successful match will be taken as one of the costs. When the crosscut is cut through a blank line, the difference between the pixel distance  $d_m$  and the standard line spacing  $d$  between the last row of the  $i$  shredder and the first row of the  $j$  shredder is taken as another cost. As a result, the cost matrix  $C2 = (c_{2ij})$  of horizontal and vertical shredding paper is obtained.

Set the variable:

$Z$ : Total cost;

$C1 = (c_{1ij})$ : Longitudinal shredding paper cost matrix;

$C2 = (c_{2ij})$ : Cost matrix of horizontal and vertical shredding paper;

$$x_{ij} = \begin{cases} 1 & \text{Assign person } i \text{ to task } j \\ 0 & \text{Do not assign person } i \text{ to task } j \end{cases}, \quad i, j = 1, 2, \dots, n$$

$$c_{1ij} = \sum \{ \text{abs}(b_k - a_k) \}, k = 1, 2, \dots, 1980;$$

$$c_{2ij} = \text{abs}(d_m - d), m = 1, 2, \dots, 1368;$$

where  $a_k$  represents the right vector of any shredded paper;  $b_k$  represents the right vector of any shredded paper;  $d_m$  represents the line spacing between the last pixel of the  $i$  shredder and the first row of the  $j$  shredder;  $d$  represents the standard line spacing.

## 2.2 Model Building

$$\min z = \sum_{j=1}^n \sum_{i=1}^n c_{ij} x_{ij} \quad (1)$$

$$\text{s.t. } \begin{cases} \sum_{i=1}^n x_{ij} = 1, & j = 1, 2, \dots, n(1) \\ \sum_{j=1}^n x_{ij} = 1, & i = 1, 2, \dots, n(2) \\ x_{ij} = 0, 1 & i, j = 1, 2, \dots, n(3) \end{cases} \quad (2)$$

where  $Z$  represents the total cost;  $c_{ij}$  represents the cost of the  $i$  individual to complete the  $j$  task;  $x_{ij}$  indicates whether the  $i$  individual is assigned to do the  $j$  task, if it is done, it is 1, and if it is not done, it is 0; constraint (1) means that each task can only be completed by one person; constraint (2) means that each person can only complete one task.

## 2.3 Nature Description

The process of solving the model is shown in Sect. 2 above, and we can find that the last updated cost matrix has the following characteristics and properties:

- All elements take only 0 or 1;
- Per row of elements have and only one 1;
- Per column of elements have and only one 1;

Property 1: if a new cost matrix is obtained by adding (minus) a constant to each element in any row (column) of the original cost matrix  $C$ , then the optimal solution to the assignment problem of the cost matrix is the same.

Therefore, using the above properties, we can transform the cost matrix to the existence of  $n$  independent zero elements (different columns in different rows), and keep each  $c_{ij}$  non-negative. At this time, let the position of the  $n$  0 elements correspond to the positions  $c_{ij} = 1$  and the other positions  $c_{ij} = 0$ , and the optimal solution can be obtained. Because it is a feasible solution with a target value of 0, and there is always a

$$z = \sum_{j=1}^n \sum_{i=1}^n c_{ij} x_{ij} \geq 0 \quad (3)$$

### 3 Problem Solving

First of all, we transform the problem of paper fragments abstractly, regard it as an assignment problem in operational research, number the pieces of paper, extract the position information on both sides of each pair of fragments to construct the location vector, and then regard the right side of one of the fragments as the task and the left side as the person who completed the task. The gray difference between two pieces of paper depends on the cost of the adult to complete the task, from which the cost matrix can be constructed, and then the new cost matrix can be obtained by the row-column transformation of the Hungarian algorithm, and the solution will be finished when the end condition of the algorithm is satisfied. At this time, we get the optimal splicing sequence of the fragments.

For the problem of paper fragments with both crosscutting and longitudinal cutting, the K-Means clustering algorithm is first used to splice in the row direction, and then the solution idea is consistent with the above description.

#### 3.1 Overview of Assignment Problem

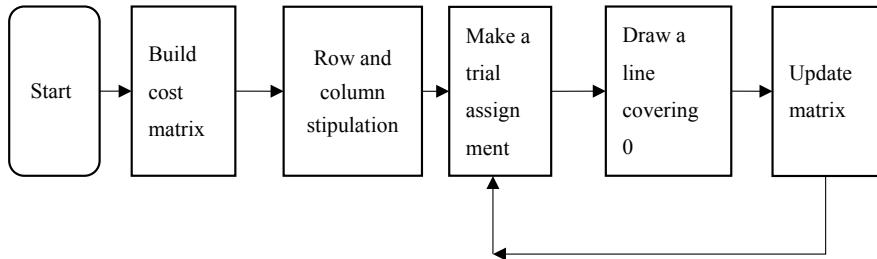
First of all, an overview of the problems solved by the Hungarian algorithm: in practice, there are n different tasks, one of which needs to be completed by n individuals, and the time for each person to complete the task is different. So there is a question of how to assign tasks to minimize time.

Generally speaking, there are n elements selected in the n \* n matrix, and there is one element in each row and column, which minimizes the sum. As shown in Table 1.

The above table can be abstracted into a matrix. If it is the minimum summation problem shown in the table above, then the matrix is called the cost matrix. If the

**Table 1** Cost matrix

Personnel/task	A	B	C	D
Personnel 1	3	12	42	15
Personnel 2	5	33	8	12
Personnel 3	35	6	13	24
Personnel 4	4	9	12	10



**Fig. 1** Flow chart of Hungarian algorithm

required problem is to maximize the sum, then the matrix is called the benefit matrix. The optimal solution of the assignment problem has such a property that if the smallest element of the row (column) of the matrix is subtracted from each element of the row (column) of the matrix, the reduced matrix is obtained, and the optimal solution is the same as that of the original matrix.

### 3.2 Hungarian Algorithm

The Hungarian algorithm is a combinatorial optimization algorithm for solving task allocation problems in polynomial time, and promotes the later original dual method. The algorithm was proposed by American mathematician Harold Kuhn in 1955. This algorithm is called the Hungarian algorithm because a large part of the algorithm is based on the previous work of the Hungarian mathematicians D énesKemonig and Jen Egerv á ry. The flow chart of the algorithm is shown in Fig. 1.

## 4 Comparison Between Experimental Results and Algorithm

### 4.1 The Solution of the Problem of Longitudinal Shredding of Paper

**Experimental setup.** An image that is all in Chinese is cut longitudinally into 19 pieces, then it is randomly scrambled and numbered. In this case, the scale of the problem is  $n = 19$ , and the cost matrix is  $19 \times 19$ . The pattern of the shredded paper is shown in Fig. 2.

The program is realized by matlab R2018a programming, and the program runs on a microcomputer with CPU2.40Ghz and 8 GB memory.



**Fig. 2** Several randomly disturbed pieces of paper

**Experimental results.** The program runs a total of 10 times, and the average running time is 0.007 s, which shows that the algorithm can obtain the minimum cost matrix and the optimal splicing sequence in a very short time. And the success rate of the splicing result obtained by running the program for 10 times is 100%. Figure 3 shows the splicing result of 19 pieces of longitudinal shredded paper, and Table 2 shows the splicing sequence and the minimum cost of 19 pieces of longitudinal shredded paper.

#### 4.2 Solution to the Problem of Horizontal and Vertical Shredding of Paper

**Experimental setup.** An image which is all in Chinese is cut longitudinally and longitudinally into  $11 * 19$  pieces, then it is randomly scrambled and numbered. First, each row is spliced successfully by using K-Means clustering algorithm, in which the scale of the problem is  $n = 11$  and the cost matrix is  $11 * 11$ . Part of the original scrap data set is shown in Fig. 4 and K-means clustering splicing of a certain line of shredded paper is shown in Fig. 5.

**Experimental results.** The program runs a total of 10 times, and the average running time is 0.008 s, which shows that the algorithm can obtain the minimum cost matrix and the optimal splicing sequence in a very short time. And the success rate of the splicing result obtained by running the program for 10 times is 100%. Figure 6 shows the splicing result of  $11 * 19$  horizontal and vertical shredded paper, and Table 3 shows the splicing sequence of  $11 * 19$  horizontal and vertical shredded paper.

#### 4.3 Performance Comparison

In the actual longitudinal shredding paper splicing scene, the most widely used method is to transform the longitudinal shredding paper problem into a traveling salesman problem based on traditional heuristic algorithms such as simulated annealing algorithm. Therefore, the Hungarian algorithm used in this paper

城上层楼叠嶂。城下清淮古汴。举手揖吴云，人与春天俱远。魂断，魂断。后夜松江月满。簌簌衣巾莎枣花。村里村北响缲车。牛衣古柳卖黄瓜。海棠珠缀一重重。清晨近帘栊。胭脂谁与匀淡，偏向脸边浓。小郑非常强记，二南依旧能诗。更有鲈鱼堪切脍，儿辈莫教知。自古相从休务日，何妨低唱微吟。天垂云重作春阴。坐中人半醉，帘外雪将深。双鬟绿坠，娇眼横波眉黛翠。妙舞蹁跹。掌上身轻意态妍。碧雾轻笼两凤，寒烟淡拂双鸦。为谁流睇不归家。错认门前过马。

我劝髯张归去好，从来自己忘情。尘心消尽道心平。江南与塞北，何处不堪行。闲离阻。谁念萦损襄王，何啻梦云雨。旧恨前欢，心事两无据。要知欲见无由，痴心犹自，情人道、一声传语。风卷珠帘自上钩。萧萧乱叶报新秋。独携纤手上高楼。临水纵横回晚粒。归来转觉情怀动。梅笛烟闻几弄。秋阴重。西山雪淡云凝冻。凭高眺远，见长空万里，云无留迹。桂魄飞来光射处，冷浸一天秋碧。玉宇琼楼，乘鸾来去，人在清凉国。江山如画，望中烟树历历。省可清言挥玉尘，真须保器全真。风流何似道家纯。不应用蜀客，惟爱卓文君。自借风流云雨散。关山有限情无限。待君重见寻芳伴。为说相思，目断西楼燕。莫恨黄花未吐，且教红粉相扶。酒阑不必看茱萸。俯仰人间今古。玉骨那愁瘴雾，冰姿自有仙风。海仙时造探芳丛。倒挂绿毛么凤。

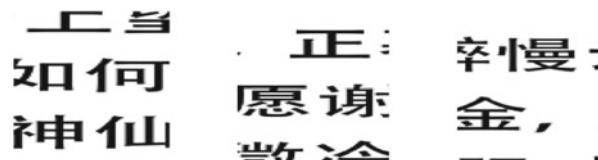
俎豆庚桑真过矣，凭君说与南荣。愿闻吴越报丰登。君王如有问，结袜赖王生。师唱谁家曲，宗风嗣阿谁。借君拍板与门檻。我也逢场作戏，莫相疑。晕腮嫌枕印。印枕嫌腮晕。闲照晚妆残。残妆晚照闲。可恨相逢能几日，不知重会是何年。茱萸仔细更重看。午夜风翻幔，三更月到床。簟纹如水玉肌凉。何物与侬归去，有残妆。金炉犹暖麝煤残。惜香更把宝钗翻。重闻处，余熏在，这一番、气味胜从前。菊暗荷枯一夜霜。新苞绿叶照林光。竹篱茅舍出青黄。霜降水痕收。浅碧鳞鳞露远洲。酒力渐消风力软，飕飕。破帽多情却恋头。烛影摇风，一枕伤春绪。归不去。凤楼何处。芳草迷归路。汤发云腴醉白，盏浮花乳轻圆。人间谁敢更争妍。斗取红窗粉面。炙手无人傍屋头。萧萧晚雨脱梧桐。谁怜季子敝貂裘。

**Fig. 3** 19 pieces of longitudinally shredded paper splicing result

**Table 2** 19 pieces of longitudinally shredded paper splicing sequence and minimum cost

Splicing sequence of 19 pieces of paper	Minimum cost value
9-15-13-16-4-11-3-17-2-5-6-10-14-19-12-8-18-1-7	65,512,725

**Fig. 4** Display of shredded paper from 11 \* 19 parts of the original horizontal and vertical cut



已属君家。且更从容等待他。愿我已无当世望，似君须向古人求。岁寒松柏肯惊秋。  
水涵空，山照市。西汉二疏乡里。新白发，旧黄金。故人恩义深。谁

**Fig. 5** The K-Means clustering splicing of a line of shredded paper

便邮。温香熟美。醉慢云鬟垂两耳。多谢春工。不是花红是玉红。一颗樱桃樊素口。不爱黄金，只爱人长久。学画鸦儿犹未就。眉尖已作伤春皱。清泪斑斑，挥断柔肠寸。嗔人问。背灯偷搵拭尽残妆粉。春事阑珊芳草歇。客里风光，又过清明节。小院黄昏人忆别。落红处处闻啼鴂。岁云暮，须早计，要褐裘。故乡归去千里，佳处辄迟留。我醉歌时君和，醉倒须君扶我，惟酒可忘忧。一任刘玄德，相对卧高楼。记取西湖西畔，正暮山好处，空翠烟霏。算诗人相得，如我与君稀。约他年、东还海道，愿谢公、雅志莫相违。西州路，不应回首，为我沾衣。料峭春风吹酒醒。微冷。山头斜照却相迎。回首向来潇洒处。归去。也无风雨也无晴。紫陌寻春去，红尘拂面来。无人不道看花回。惟见石榴新蕊、一枝开。

九十日春都过了，贪忙何处追游。三分春色一分愁。雨翻榆荚阵，风转柳花球。白雪清词出坐间。爱君才器两俱全。异乡风景却依然。团扇只堪题往事，新丝那解系行人。酒阑滋味似残春。

缺月向人舒窈窕，三星当户照绸缪。香生雾縠见纤柔。搔首赋归欤。自觉功名懒更疏。若问使君才与术，何如。占得人间一味愚。海东头，山尽处。自古空槎来去。槎有信，赴秋期。使君行不归，别酒劝君君一醉。清润潘郎，又是何郎婿。记取钗头新利市。莫将分付东邻子。西塞山边白鹭飞。散花洲外片帆微。桃花流水鳜鱼肥。主人瞋小。欲向东风先醉倒。已属君家。且更从容等待他。愿我已无当世望，似君须向古人求。岁寒松柏肯惊秋。

水涵空，山照市。西汉二疏乡里。新白发，旧黄金。故人恩义深。谁道东阳都瘦损，凝然点漆精神。瑤林终自隔风尘。试看披鹤氅，仍是谪仙人。三过平山堂下，半生弹指声中。十年不见老仙翁。壁上龙蛇飞动。暖风不解留花住。片片著人无数。楼上望春归去。芳草迷归路。犀钱玉果。利市平分沾四坐。多谢无功。此事如何到得侬。元宵似是欢游好。何况公

卿早入小窗中，共赏此景。不知君意何如？

**Fig. 6** 11 \* 19 pieces of horizontal and vertical shredded paper splicing result

**Table 3** 11 \* 19 pieces of horizontal and vertical shredded paper splicing sequence

Splicing sequence of 11 pieces of horizontal shredded paper
10-11-8-7-2-9-4-1-6-3-5

is compared with the simulated annealing algorithm represented by the traditional heuristic algorithm, and the splicing time, splicing success rate and time complexity of the algorithm are compared respectively.

For the problem of longitudinal shredding of paper, the simulated annealing algorithm program was run for 10 times, and the average running time was 131.876 s. The success rate of the splicing result obtained by running the program for 10 times is 100%, and the splicing result is the same as that in Fig. 3. Table 4 shows the splicing sequence of 19 pieces of longitudinally shredded paper.

For the problem of horizontal and vertical shredding of paper, the simulated annealing algorithm program was run for 10 times, and the average running time

**Table 4** Simulated annealing algorithm for solving the splicing sequence of 19 pieces of longitudinally shredded paper

Splicing sequence of 19 pieces of longitudinally shredded paper
16–4–11–3–17–2–5–6–10–14–19–12–8–18–1–7–9–15–13

**Table 5** Simulated annealing algorithm for solving the splicing sequence of 11–19 pieces of horizontal and longitudinal shredded paper

Splicing sequence of 11 to 19 pieces of horizontal and vertical shredded paper
6–2–7–11–10–5–8–3–4–1–9

was 161.147 s. The success rate of the splicing result of the shredded paper obtained by running the program for 10 times is 100%, and the splicing result of the shredded paper is the same as that of Fig. 6. Table 5 shows the splicing sequence of 11–19 horizontal and vertical shredded paper.

As can be seen from Tables 6 and 7, the bold data represent the fastest splicing time and minimum time complexity of the two algorithms. From the analysis of the time complexity of the algorithm itself, the time complexity of the Hungarian algorithm used in this paper is  $O(n^3)$ , and the time complexity of simulated annealing algorithm is  $O(n^n)$ . We can find that when the problem size  $n \geq 3$ , the time complexity of the Hungarian algorithm is better than that of the simulated annealing algorithm. Under the premise that the scale of the longitudinal shredding problem set in this

**Table 6** Comparison of splicing time and time complexity of different algorithms for longitudinal shredding problem

Arithmetic	Problem scale	Splicing takes time	Time complexity	Splicing success rate
Simulated annealing algorithm	19	131.876 s	$O(n^n)$	100%
Hungarian algorithm	19	<b>0.007 s</b>	<b><math>O(n^3)</math></b>	100%

**Table 7** Comparison of splicing time and time complexity of different algorithms for horizontal and vertical shredding problem

Arithmetic	Problem scale	Splicing takes time	Time complexity	Splicing success rate
Simulated annealing algorithm	11	161.147 s	$O(n^n)$	100%
Hungarian algorithm	11	<b>0.008 s</b>	<b><math>O(n^3)</math></b>	100%

paper is  $n = 19$ , the problem scale of the horizontal and longitudinal shredding problem after K-Means clustering is  $n = 11$ , and the success rate of 100% shredded paper splicing is guaranteed at the same time, the comparison of the solution time of the two algorithms is the most intuitive embodiment. The two algorithms also run 10 times, in the longitudinal shredding problem, the average running time of the Hungarian algorithm is 0.007 s, the average running time of the simulated annealing algorithm is 131.876 s; in the horizontal and vertical shredding problem, the average running time of the Hungarian algorithm is 0.026 s, the average running time of the simulated annealing algorithm is 161.147 s. It can be seen that the average running time of the Hungarian algorithm is much lower than that of the simulated annealing algorithm, and the gap between the two solving time will be geometrically enlarged with the expansion of the scale of the problem. Thus it can be seen that the Hungarian algorithm used in this paper not only greatly reduces the amount of time and code running time in the process of splicing, but also ensures the success rate of splicing.

## 5 Conclusion

In view of the problems of traditional heuristic algorithm in splicing longitudinal shredded paper, such as large amount of calculation, low success rate and narrow application scene, this paper changes the idea of the problem and transforms the problem into an assignment problem in operational research. The Hungarian algorithm is used to solve the problem. The experimental results show that the Hungarian algorithm used in this paper is better than the simulated annealing algorithm in splicing time and time complexity, and can ensure the splicing success rate. Therefore, the algorithm used in this paper is feasible and efficient.

The Hungarian algorithm used in this paper mainly needs to be improved: when a Chinese image and an English image are longitudinally cut at the same time and randomly disturbed together, it is not advisable to directly use the Euclidean distance calculated by the position vectors on the left and right sides of two random pieces of paper as the elements of the cost matrix, and it is very likely that all two different images will be spliced together. In addition, when the graphics fragments are irregular, it will increase the complexity of the construction cost matrix, at the same time increase the difficulty of splicing, and increase the splicing time and splicing error rate. Therefore, based on the above problems, in the face of more complex shredding problems, the rule matching algorithm will be added to improve the solution efficiency and splicing success rate of the whole algorithm.

**Acknowledgements** Fund project: Hunan Province Key Research And Development Plan Project: 2018GK2058, Hunan Provincial Natural Science Foundation of China: 2020JJ4248, 2020JJ4251, 2018JJ3264, A Project Supported by Scientific Research Fund of Hunan Provincial Education Department: 20B142, Degree & Postgraduate Education Reform Project of Hunan Province: 2019JGYB242.

## References

1. Du, H., Liu, X., Ca, L.: Splicing and restoration of large quantities of single-sided scraps based on improved genetic algorithm. *Electron. World* **13**, 52–53 (2018)
2. Qi, Y., Yang, X., Wei, M., Yun, L., Wang, Q., Ping, A.N.: Mosaic restoration of paper fragments based on clustering analysis and Euclidean distance model. *Electron. Technol. Softw. Eng.* **18**, 145–146 (2020)
3. Mao, X.: Restoration model of shredded paper splicing based on linear programming. *Inf. Recording Mater.* **19**(10), 209–210 (2018)
4. Yu, C., Chen, J.T.: A method of splicing and restoring equal size rectangular scraps of paper. *J. Xiamen Univ. Technol.* **22**(3), 103–108 (2014)
5. Mu, H., Xv, Y., Yi, Q., Ying Qun., Li, J., Li, P.: A method of stitching irregular scrapped paper base on edge position and color information. *J. Shaanxi Normal Univ. (Nat. Sci. Ed.)* **48**(6), 90–95 (2020)
6. Dang, Y.C., Wan, L., Zhou, Q.: Research on shredded paper splicing reconstruction algorithm based on evolutionary computation. *Res. Explor. Lab.* **48**(6), 131–136+214 (2020)
7. Lou, Z.Z.: Semi-auto stitching of scrapped paper based on character characteristic. *Comput. Eng. Appl.* **48**(5), 207–210 (2012)
8. Zhuang, S.F., Fu, X.M.: A 0–1 integer programming based restoration algorithm for shredded paper. *J. Shaoguan Univ. Nat. Sci.* **38**(9), 9–14 (2017)
9. Liu, Q.J., Chen, P., Wang, Z.Y.: Algorithm design on scraps of paper splicing based on text feature. *Res. Explor. Lab.* **35**(11), 110–113 (2016)
10. Fu, J.H., Hua, Y., Chen, J.H., Pan, H.W.: Horizontal and vertical shredded paper restoration algorithm based on clustering and ant colony algorithm. *Math. Pract. Theor.* **49**(15), 199–209 (2019)
11. Han, Y.Y., Zhang, Y.P., Shen, H.P., Wang, Y.K.: Rule graph fragment mosaic based on genetic algorithm and 0–1 programming. *Electron. Sci. Tech.* **28**(5), 136–139 (2015)
12. Yang, L., He, L.: Automatic 0–1 programming based reassembly of fragmented chinese documents. *Comput. Syst. Appl.* **24**(4), 270–273 (2015)
13. Ding, Z.Q.: The semi-automatic restoration of regular paper fragments. *Appl. Mech. Mater.* **3335** (2014)
14. Zhou, A.J.: The auto restoration of paper fragments rules based on the traveling salesman model. *Appl. Mech. Mater.* **3056** (2014)
15. Liang, X., Qin, H., Shi, J., Huang, L., Jiatai, G.: A study of automatically assembling paper fragments on simulated annealing algorithm [P]. In: Proceedings of the 2015 International Conference on Electrical, Computer Engineering and Electronics (2015)

# Rendezvous Control for Autonomous Underwater Vehicles with Event Triggered Cloud Access



Feng Zhou , Ge Zheng , and Kewu Tao 

**Abstract** This paper investigates a multi-agent consensus control problem with event triggered communication for underwater robots. The underwater robots can only send and receive information through an asynchronous intermittently connection with cloud repository when they surface to the water for the reason that the underwater communication devices are expensive. To minimize the amount that robot surfaces, a novel self-triggered control strategy that ascertain when the communication with cloud is needful is presented to guarantee the global consensus task. Different with previous event triggered control, the underwater robot can't receive the newest triggering information of neighbors until the next time it surfaces. A useful theorem is obtained to prove that the entire system removes the Zeno behavior and also consensus building in the end. Simulation results show that the proposed self-trigger consensus policy is correct and reliable.

**Keywords** Event-triggered control · Autonomous underwater vehicles · Rendezvous

## 1 Introduction

During the practice, the application of autonomous underwater vehicles has captured wide notice, such as ocean exploration and patrol. Multiple autonomous underwater vehicles have made many accomplishments, see as [1–3]. Most of these papers use expensive sonar modems to achieve underwater communication. It is well known that the communication underwater is indeed particularly severe [3, 4], the radio communication is blocked. Even though the communication transmits through sonar, the job range and frequency bandwidth will also be limited. Moreover, the GPS signal is not available underwater, the accurate positions of robots are hard to adopt in the sea.

---

F. Zhou  · G. Zheng · K. Tao

School of Electrical and Information Engineering, Changsha University of Science and Engineering, Changsha 410075, Hunan, China

e-mail: [zhoufengcsu@csust.edu.cn](mailto:zhoufengcsu@csust.edu.cn)

Motivated by this fact, [5] proposed a periodic strategy for underwater agents to surface and exchange information with each other on the water through wireless communication in which the expensive sonar modems are no longer essential. However, this strategy requires all the underwater agents to surface at the same time. In order to relax this requirement, [6, 7] introduce a cloud access for the underwater multi-agent systems to support a shared asynchronous communication database. Thus, the information can be stored in the cloud, and the underwater agents can receive and send the information intermittently on the cloud in an asynchronous way. The remaining also the most important work for this kind of scenario is how to schedule the time sequence that each agent needs to surface to communicate with cloud.

The method used in these papers is event-triggered control, which is designed to tune the time intervals of control updates and communication transmission. If a control signal update indicates that a specific event has occurred, which is the only feature, [8, 9]. This incident is defined by making a tradeoff between resource consumption (i.e. computation, communication, actuator efforts) and performance. In [10], a criterion is presented to show how to determine the time that the control signal should be updated. In [11], the problem of output feedback stabilization of continuous-time networked systems with persistent external disturbances is solved. To further resource consumption, in [12, 13], a kind of new event trigger synchronization method is proposed, which combines event trigger control with periodic sampling date control. Self-trigger control is based on event-trigger control in which the next trigger time is computed based on the event condition and last trigger time [14–16]. Thus the requirement of continuous verification that whether the event condition is triggered is relaxed.

Event-triggered control is also widely used in consensus research on multi-agent systems. For example, the work [17] proposed a time-dependent event-triggered strategy in which the event condition uses only its own information to judge whether it should be triggered. In [18] a centralized and decentralized consensus event-triggered consensus strategy is proposed, its state-dependent threshold composed of local information, the drawback of this strategy is that it required each agent has perfect access to information of neighbors. Based on this method, [19] proposed a similarity strategy where the requirement of continuous information of neighbors is relaxed. In [20], a combinational measurement approach is proposed to event design, the continuous measurement of neighbors is also avoided.

However, all of these previous works are assumed that all the agents can receive the information of neighbors immediately once the event is triggered by some neighbor agent. This assumption is not applicable for underwater robots scenarios in this paper since we consider the underwater agent can receive the information only when its own event is triggered (i.e. it surfaces). Therefore, in this paper, we are interested in developing a novel event-triggered strategy for underwater robots scenario. Our work is similar to [6, 7], while [6] proposed an event triggered solution for underwater agents with noise, in which the event condition is time-dependent and the system finally reaches a ball range around the desired consensus state. In [7], a

state-dependent self-triggered strategy is proposed to better schedule the events with global consensus, but the possibility of Zeno behavior can't be removal.

In this study, the distributed event triggered control is investigated for underwater robots to reduce the amount that robot surfaces. The strategy doesn't require the agents can be aware of whether the neighboring agents are triggered and receive the newest triggering state of neighbors immediately. Each agent schedules its own sequence of time when it needs to surface to achieve the global task. The contributions of this paper are twofold: first, a novel simply self-triggered consensus strategy is proposed for underwater robots. Second, it is strictly proved in theory that guarantees there is no Zeno behavior the overall system will achieve the consensus.

The rest of the chapters are as follows. Some symbols and problem descriptions are introduced in Sect. 2. In Sect. 3, a detailed event triggered consensus control design method is presented, and it is shown how this strategy is attained by scheduling the control updates in a self-triggered way. Section 4 shows the simulation results of the proposed strategy. Section 7 concludes the paper with some possible future developments.

## 2 Preliminaries and Problem Statement

This section introduces the related theories of graph theory. Let  $G(V, E, A)$  denote a weighted directed graph, where  $V = \{1, 2, \dots, N\}$  is means there are  $N$  agents and that's the set of all agents and  $E = \{(i, j) \in V \times V\}$  is the set of edges between nodes,  $A = [a_{ij}]$  is the weighted adjacency matrix. An edge of  $G$  is  $(i, j) \in E$ , then  $a_{ij} > 0$ . For undirected graph,  $a_{ij} = a_{ji}$ . The set of neighbors of the node  $i$  is denoted by  $N_i = \{j \in V : (i, j) \in E\}$ . The degree matrix of graph  $G$  is denoted by  $D = diag\{d_1, d_2, \dots, d_N\}$ , where  $d_i = \sum a_{ij}$ , then the Laplacian matrix  $L$  is defined as  $L = D - A$ . For connected graphs,  $L$  has exactly one zero eigenvalue  $\lambda_1(G)$  and the smallest non-zero eigenvalue  $\lambda_2(G)$  is called algebraic connectivity. Suppose the topology in this paper is a fixed undirected graph.

This paper studies a multi-agent system composed of  $N$  multi-agents based on first-order dynamic model, described as

$$\dot{x}_i(t) = u_i(t) \quad (1)$$

where  $x_i$  denotes the state of agent  $i$  (i.e. 3D position),  $u_i$  denotes the control input for each agent. The consensus problem of underwater vehicles is studied in this paper, such that  $x_i(t) - x_j(t) \rightarrow 0$  with  $t \rightarrow \infty$  for all  $i, j \in \{1, \dots, N\}$ .

It is assumed that each agent cannot communicate directly with other agents in water, because the radio communication is blocked underwater, and expensive sonar modems aren't equipped on the robots. All the agents should surface to the water at discrete time instants to connect a shared cloud to upload and download information, such that the agents can have access to the information of neighbors asynchronously.

Let  $\{t_i^k\}$  be the sequence of time that each agent surfaces.  $x\{t_i^k\}, u\{t_i^k\}$  are defined as the current state and control input respectively responding to  $\{t_i^k\}$ . When agent  $i$  surfaces to connect to the cloud, it has to do three tasks: download all the information of its neighbors such as the last time its neighbors surfaced  $\{t_j^{last}\}$ , last updated state of its neighbors  $u_j\{t_j^{last}\}$ , and last updated control input  $u_j\{t_j^{last}\}$ . Using the received information of its neighbors to compute the new control input  $u_i\{t_i^{last}\}$ . Upload the information of itself such as current time it surfaces  $\{t_i^{last}\}$ , current state  $x_i\{t_i^{last}\}$  and the updated control input  $u_i\{t_i^{last}\}$ . Obviously,  $t_i^{last} = t_i^k$  for any given  $t \in [t_i^k, t_i^{k+1}]$ . For simplicity, it is assumed that the communication with the cloud is instantaneous, there is no communication delay.

Since the communication underwater is assumed completely disconnected, the common distributed continuous control law  $u_i = - \sum_{j \in N_i} a_{ij}(x_i - x_j)$  is unavailable in this scenario, which requires each agent to have access to continuous information of its neighbors. Moreover, the recent event triggered consensus control law is also can't be directly applied to underwater robots due to the drawback that all agents are required to receive the information of neighbors immediately. In other words, the neighbors are needed to be immediately aware of whether the agent  $i$  is triggered and can update its control law accordingly. So the purpose of this paper is to develop a novel event triggered consensus strategy to determine the sequence of time that each agent surfaces such that the underwater system converges to consensus gradually.

### 3 Distributed Event Triggered Rendezvous Design

With the analysis above, control signal is updated only when an event is triggered (i.e., when the agent surfaces). So the control input is a piecewise constant function with event triggered updates

$$u_i(t) = u_i(t_i^k) = \sum_{j=1}^N a_{ij}(x_j(t_i^k) - x_i(t_i^k)) \quad (2)$$

for  $t \in [t_i^k, t_i^{k+1})$ ,  $x_j\{t_i^k\}$  is the state of agent  $j$  at time  $t_i^k$  when agent  $i$  surfaces, which can't be obtained directly from cloud. However, agent  $i$  can use the information available in the cloud to calculate the state of adjacent agents.

$$x_j(t_i^k) = x_j(t_j^{last}) + u_j(t_j^{last})(t_i^k - t_j^{last}) \quad (3)$$

For the purpose of introducing event triggered strategy, we define a mismatch error  $e_i(t)$  to every agent as

$$e_i(t) = u_i(t) - z_i(t) \quad (4)$$

where

$$z_i(t) = \sum_{j=1}^N a_{ij} (x_j(t) - x_i(t)) \quad (5)$$

is the common distributed continuous control law which also describes the state error among each agent. When an event is triggered by agent  $i$  we will have  $e_i(t_i^k) = u_i(t_i^k) - z_i(t_i^k) = 0$ , because  $t = t_i^k$  is an event triggering time for agent  $i$ .

In the event design, design event condition is the essential task. If the condition is triggered, an event occurs and the controller updates. The design method for event condition is described in the following part and a lemma is derived.

**Lemma 1** *Assuming that the communication graph is a connected undirected graph, the multi-agent system (1) under control (2) is globally gradually consistent with the event condition*

$$\|e_i(t)\| \leq \sqrt{\sigma_i(2a - a^2)} \|u_i(t)\| \quad (6)$$

for  $0 < \sigma_i < 1$  and  $0 < a < 2$ .

**Proof** For system (1) with trigger condition (6), we choose the Lyapunov function as

$$V(t) = \frac{1}{2} x^T(t) L x(t) \quad (7)$$

Taking the derivative of  $V(t)$  along the trajectory (2) yields

$$\begin{aligned} \dot{V}(t) &= x^T(t) L \dot{x}(t) = -x^T(t) L u(t) = -z^T(t) u(t) = -(u(t) - e(t))^T u(t) \\ &= -\sum_i^N \|u_i(t)\|^2 + \sum_i^N e_i^T(t) u_i(t) \end{aligned} \quad (8)$$

For any given  $a > 0$ ,  $|xy| \leq \frac{a}{2}x^2 + \frac{1}{2a}y^2$  holds for undirected graph. Thus we can get

$$\begin{aligned} \dot{V}(t) &\leq -\sum_i^N \|u_i(t)\|^2 + \sum_i^N \frac{a}{2} \|u_i(t)\|^2 + \sum_i^N \frac{1}{2a} \|e_i(t)\|^2 \\ &= -\sum_i^N \left(1 - \frac{a}{2}\right) \|u_i(t)\|^2 + \sum_i^N \frac{1}{2a} \|e_i(t)\|^2 \end{aligned} \quad (9)$$

Applying the event condition (6), we will have

$$\dot{V}(t) \leq \sum_i^N (\sigma_i - 1) \left(1 - \frac{a}{2}\right) \|u_i(t)\|^2 \quad (10)$$

Which implies  $\dot{V}(t) \leq 0$  for  $0 < \sigma_i < 1$  and  $0 < a < 2$ .

Due to  $V \geq 0$ ,  $\dot{V} \leq 0$  that means  $V$  has a finite restraint and  $\dot{V} \rightarrow 0$  as  $t \rightarrow \infty$ . Then we can get

$$0 = \lim_{t \rightarrow \infty} \dot{V}(t) \leq \sum_i^N (\sigma_i - 1) \left(1 - \frac{a}{2}\right) \|u_i(t)\|^2 \leq 0 \quad (11)$$

Thus, from Eq. (4), we can get  $\lim_{t \rightarrow \infty} u_i(t) = 0$ ,  $\lim_{t \rightarrow \infty} e_i(t) = 0$ .

In the meanwhile,

$$0 = \lim_{t \rightarrow \infty} \dot{V}(t) = \lim_{t \rightarrow \infty} (-z(t)^T z(t) - z(t)^T e(t)) \quad (12)$$

Because  $\lim_{t \rightarrow \infty} e_i(t) = 0$ . The equation can be rewritten as

$$\lim_{t \rightarrow \infty} \dot{V}(t) = \lim_{t \rightarrow \infty} (-z(t)^T z(t)) = -\lim_{t \rightarrow \infty} \sum_i^N \|z_i(t)\|^2 = 0 \quad (13)$$

That is  $\lim_{t \rightarrow \infty} z_i(t) = 0$  for  $i = 1, 2, \dots, n$ . According to the definition of  $z_i(t)$ , we have  $\lim_{t \rightarrow \infty} \sum_{j=1}^N a_{ij} (x_j(t) - x_i(t)) = 0$ . Therefore  $\lim_{t \rightarrow \infty} x_i(t) = \lim_{t \rightarrow \infty} x_j(t)$ ,  $i, j = 1, 2, \dots, n$ . Each agent comes to a consensus building over time in the end.

It should be special noted that the event condition (6) requires exact neighbors' state of each agent which can be computed as

$$x_j(t) = x_j(t_j^{last}) + u_j(t_j^{last})(t - t_j^{last}) \quad (14)$$

But this equation is only valid when  $t \leq t_j^{next}$ ,  $t_j^{next}$  is the next trigger time for agent  $j$  after  $t_j^{last}$ . If agent  $j$  resurfaces and updates its control input during the time interval  $(t_i^{last}, t_i^{next})$ , the Eq. (11) is no longer correct. So in the following, we propose a self-triggered control to schedule and implement the event triggered control updates.

From (4) one has  $\dot{e}_i(t) = -\dot{z}_i(t)$ . Thus in the interval  $(t_i^k, t_i^{k+1})$  we can get

$$\frac{d}{dt} \|e_i(t)\| \leq \|\dot{z}_i(t)\| = \left\| \sum_{j=1}^N a_{ij} (\dot{x}_j(t) - \dot{x}_i(t)) \right\| \leq \left\| \sum_{j=1}^N a_{ij} u_j(t_i^k) \right\| + \left\| \sum_{j=1}^N a_{ij} u_j(t_j^{k'}) \right\| \quad (15)$$

where  $t_j^{k'}$  is the last triggering time of agent  $j$  before  $t$ . Observe that

$$\|u_i(t)\| \leq \|e_i(t)\| + \|z_i(t)\| \leq \sqrt{\sigma_i(2a - a^2)}\|u_i(t)\| + \|z_i(t)\| \quad (16)$$

we can get  $\left(1 - \sqrt{\sigma_i(2a - a^2)}\right)u_i(t) \leq z_i(t)$ , then the Eq. (10) could be rewritten as

$$\dot{V}(t) \leq \sum_i^N (\sigma_i - 1) \left(1 - \frac{a}{2}\right) \|u_i(t)\|^2 \leq -\sum_i^N \frac{(\sigma_i - 1)(1 - \frac{a}{2})}{\left(1 - \sqrt{\sigma_i(2a - a^2)}\right)} \|z_i(t)\|^2 \leq -\sigma \lambda_n V(t) \quad (17)$$

where  $\sigma = \max_{i=1,\dots,n} \frac{(\sigma_i - 1)(1 - \frac{a}{2})}{\left(1 - \sqrt{\sigma_i(2a - a^2)}\right)}$ , Thus we have  $V(t) \leq V(t_1)e^{-\sigma \lambda_n(t-t_1)}$ .since  $V(t)$  is non-increasing, one can obtain that for each agent  $u_i(t) = z_i(t_i^k) \leq \sqrt{\lambda_n V(t_i^k)} \leq \sqrt{\lambda_n V(0)e^{-\sigma \lambda_n t_i^k}}$ . Then the Eq. (12) can be rewritten as

$$\frac{d}{dt} \|e_i(t)\| \leq d_i \|u_i(t_i^k)\| + d_i \sqrt{\lambda_n V(0)e^{-\sigma \lambda_n t_{\min}}} \quad (18)$$

where  $d_i = \sum a_{ij}$  is denoted in the notation,  $t_{\min} = \min_{j \in N_i} t_j^{k'}$  is the smallest time among the last triggering time of neighbor agents.

So the time when  $e_i(t)$  evolves from 0 to  $\sqrt{\sigma_i(2a - a^2)}u_i(t)$  is when the next event is triggered, which could be lower bounded by

$$t_i^{k+1} = t_i^k + \frac{\sqrt{\sigma_i(2a - a^2)}\|u_i(t_i^k)\|}{d_i \|u_i(t_i^k)\| + d_i \sqrt{\lambda_n V(0)e^{-\sigma \lambda_n t_{\min}}}} \quad (19)$$

Consequently, the interval between two triggering time is a strictly positive value. the updates do not present accumulation points, that means Zeno behavior have been ruled out.

Obviously, all the needed information by (16) is available or could be computed through (7) and (11). Thus the event-triggered strategy could be realized for each agent. From the analysis above we can conclude that the systems will consensus with the sequence of self-trigger times  $\{t_i^k\}$  determined by (16). However, directly using (16) is simply but rather conservative, since the control law of agent  $j$  may change if agent  $j$  is triggered in the interval  $(t_i^k, t_i^{k+1})$ , then the upper bound of  $\|u_j(t_j^{k'})\|$  will change.

We define  $t'_{\min} = \min_{j \in N_i} t_j^{next}$  is the smallest time among the next triggering time of neighbor agents, and  $\tau_i^k$  is the next triggering time computed by event condition (6) and neighbor trajectory (11). It is easily to see that the event condition (6) is not satisfied when  $t > \tau_i^k$ . Then if  $\tau_i^k \leq t'_{\min}$ , which means there is no neighbor triggered when agent  $i$  dive in the water, thus the next triggering time could be easily set as

$$t_i^{k+1} = \tau_i^k \quad (20)$$

Otherwise, one or more neighbor agent will be triggered before its next surfacing. The control laws and trajectories of neighbors will be changed when they are triggered. Otherwise, one or more neighbor agent will be triggered before its next surfacing. The control laws and trajectories of neighbors will be changed when they are triggered. So for  $[t_i^k, t'_{\min}]$ , the information of neighbors are still available

$$\frac{d}{dt} \|e_i(t)\| \leq \|\dot{z}_i(t)\| = \sum_{j=1}^N a_{ij} \|u_i(t_i^k) - u_j(t_j^{last})\| \quad (21)$$

For  $[t'_{\min}, t_i^{k+1}]$ , the control law of triggering neighbors are no longer known, so similar with (15), the evolution of  $\|e_i(t)\|$  could be bound as

$$\begin{aligned} \frac{d}{dt} \|e_i(t)\| \leq \|\dot{z}_i(t)\| &= \sum_{j=1}^N a_{ij} \|u_i(t_i^k) - u_j(t_j^{next})\| \\ &\leq d_i \|u_i(t_i^k)\| + d_i \sqrt{\lambda_n V(0) e^{-\sigma \lambda_n t'_{\min}}} \end{aligned} \quad (22)$$

Then the next trigger time for agent  $i$  could be lower bounded and computed as

$$t_i^{k+1} = t'_{\min} + \frac{\sqrt{\sigma_i(2a - a^2)} \|u_i(t_i^k)\|}{d_i \|u_i(t_i^k)\| + d_i \sqrt{\lambda_n V(0) e^{-\sigma \lambda_n t'_{\min}}}} - \frac{(t'_{\min} - t_i^k) \sum_{j=1}^N a_{ij} \|u_i(t_i^k) - u_j(t_j^{last})\|}{d_i \|u_i(t_i^k)\| + d_i \sqrt{\lambda_n V(0) e^{-\sigma \lambda_n t'_{\min}}}} \quad (23)$$

**Remark 1** Note that in Eq. (19), the next trigger time  $t_j^{next}$  is must known. Thus when each agent surfaces to the water, it must do more works such as compute the next triggering time of itself and upload to the cloud and download the next triggering time of neighbors. Specifically, if some agent  $j$  and  $i$  are surfacing at the same time, the next trigger time  $t_j^{next}$  is still can be computed by (20). At this situation, we have  $t'_{\min} = \min_{j \in N_i} t_j^{next} = t_i^k$ , because the  $t_j^{next}$  downloaded from the cloud has not updated yet.

Then the formal self-triggered rendezvous algorithm can be concluded in the following Algorithm 1 and the main results are summarized in Theorem 1.

**Algorithm 1** Self-triggered rendezvous algorithm

---

At every triggering time  $t_i^k$ , agent  $i \in \{1, \dots, n\}$  will do:

---

1: download  $t_j^{last}$ ,  $x_j(t_j^{last})$ ,  $u_j(t_j^{last})$ ,  $t_j^{next}$  for all neighbors  $j \in N_i$  from cloud;

2: compute the state of neighbors at time  $t_i^k$ ,  $x_j(t_i^k) = x_j(t_j^{last}) + u_j(t_j^{last})(t_i^k - t_j^{last})$ ;

3: compute the updated control input,  
 $u_i(t_i^k) \sum_{j=1}^N a_{ij}(x_j(t_i^k) - x_i(t_i^k))$ ;

4: compute for the  $\tau_i^k$  by using event condition (6) and (11);

5: compute  $t'_{min} = \min_{j \in N_i} t_j^{next}$ ;

6: **if**  $\tau_i^k \leq t'_{min}$  **then** set  $t_i^{k+1} = \tau_i^k$ ;

7: **else**

8: compute the next triggering time  $t_i^{k+1}$  by (20);

9: **End if**

10: upload the state information of itself,  $t_i^{last} = t_i^k$ ,  $x_i(t_i^{last})$ ,  $u_i(t_i^{last})$ ,  $t_i^{next} = t_i^{k+1}$ ;

11: dive into the water and run using the control law  
 $u_i(t) = u_i(t_i^k)$  for  $t \in [t_i^k, t_i^{k+1}]$ , surface at time  $t_i^{k+1}$

---

**Theorem 1** Consider the multiple autonomous underwater vehicles (1) with the control law (2). Suppose that the communication graph is connected and undirected. Then the underwater vehicle systems (1) are globally asymptotically rendezvous with the triggering time sequence  $\{t_i^k\}$  determined by Algorithm 1 for all agent  $i \in \{1, \dots, n\}$ .

## 4 Simulations

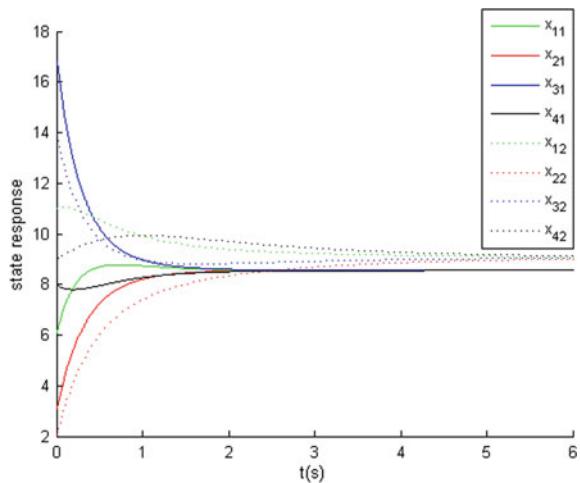
Consider there are 4 agents in the system and apply the control design method described in Sect. 3 to the system.

The initial value of each agent is chosen as  $x_1(0) = [6; 11]$ ,  $x_2(0) = [3; 2]$ ,  $x_3(0) = [17; 14]$  and  $x_4(0) = [8; 9]$ . Choose a set of parameters  $a = 1$ ,  $\sigma = 0.3$ . the Laplacian matrix is given by

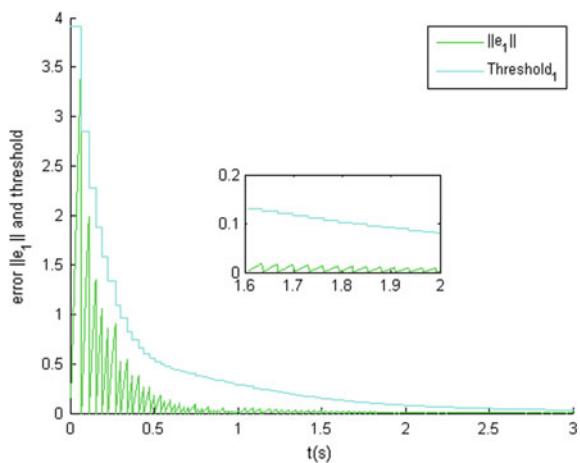
$$L = \begin{bmatrix} 2 & 0 & -1 & -1 \\ 0 & 1 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \quad (24)$$

Figure 1, 2, 3, 4, 5 shows the simulation results. Figure 1 describes the state trajectory of four agents under proposed self-triggered control. It can be seen from the figure that the four agents converge gradually over time, indicating that the proposed design method is correct. Figures 2, 3, 4 and 5 shows the evolution of the mismatch error  $\|e_1\|$  for each agent under proposed self-triggered control respectively. From these 4 figure, we can see how the framework is realized in the distributed case for each agent. Once an event is triggered at all hours, the error signal is reset to zero immediately, and the control input is updated to the real value and held during the next time interval. Thus the threshold is also a piecewise constant function. Specifically, in Figs. 2, 3, 4 and 5, it can be seen that the event is triggered even though the error signal is not surpassed the state-dependent threshold. This is because the time defined

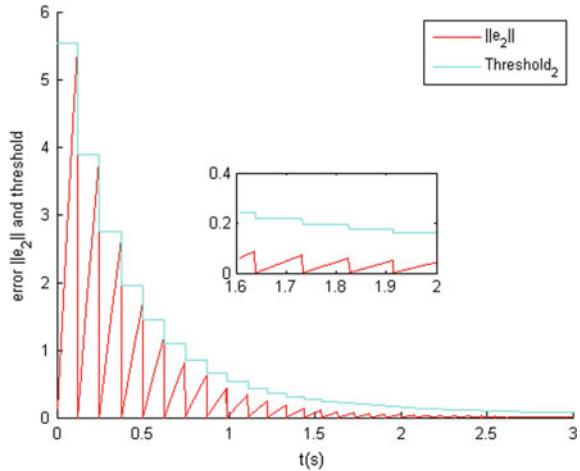
**Fig. 1** State trajectory of four agents under proposed self-triggered control



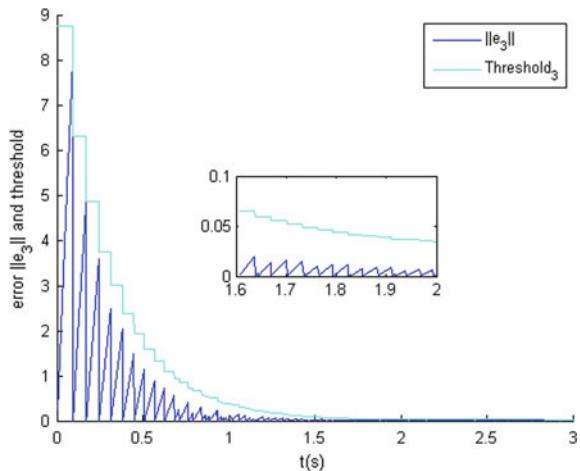
**Fig. 2** The threshold for node 1 and the evolution of mismatch error  $\|e_1\|$



**Fig. 3** The threshold for node 2 and the evolution of mismatch error  $\|e_2\|$



**Fig. 4** The threshold for node 3 and the evolution of mismatch error  $\|e_3\|$

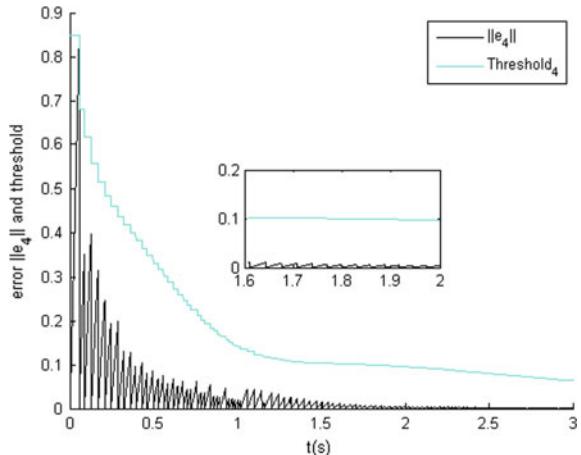


by Algorithm 1 is a little conservative. With the agents converging to rendezvous, the state-depend threshold tends to zero. From the above analysis, it can be concluded that the rendezvous control of the underwater vehicle can show good convergence performance under the event trigger control.

## 5 Conclusion

In this paper, a novel self-triggered control strategy is proposed for consensus of a network of underwater robots with single-integrator dynamics, in which the agents

**Fig. 5** The threshold for node 4 and the evolution of mismatch error  $\|e_4\|$



can only communicate with each other via an asynchronous intermittently communication with a cloud. The strategy doesn't require the agents can be aware of whether the neighboring agents are triggered and receive the newest triggering state of neighbors immediately. Each agent schedules its own sequence of time when it need to surface to achieve the global task. Specifically, this strategy is pragmatic and strictly proved in theory that guarantees there is no Zeno behavior the overall system will achieve the consensus. Future work will be carried out to apply the proposed strategy to other complex models or multi-agent systems with disturbance.

**Acknowledgements** This work was supported by Changsha Municipal Natural Science Foundation (Grant No. 2021cskj014), the Project Funded by the Hunan Provincial Department of Education under Award 18C0203.

## References

1. Gun-Rae, C., Jihong, L., Daegil, P.: Robust trajectory tracking of autonomous underwater vehicles using back-stepping control and time delay estimation. *Ocean Eng.* **201**, 107131 (2020)
2. Yufei, Z., Haibing, H., Sharma, S., Dianguo, X., Qiang, Z.: Cooperative path planning of multiple autonomous underwater vehicles operating in dynamic ocean environment. *ISA Trans.* **94**, 174–186 (2019)
3. Tingting, Y., Shuanghe, Y., Yan, Y.: Formation control of multiple underwater vehicles subject to communication faults and uncertainties. *Appl. Ocean Res.* **82**, 109–116 (2019)
4. Suryendu, C., Subudhi, B.: formation control of multiple autonomous underwater vehicles under communication delays. *IEEE Trans Circ. Syst. II: Express Briefs* **99**, 1–1 (2020)
5. Teixeira, P., Dimarogonas, D., Johansson, K., Sousa, J.: Multi-agent coordination with event-based communication. In: American Control Conference, pp. 824–829 (2010)
6. Adaldo, A., Liuzza, D., Dimarogonas, D., Johansson, K.: Control of multi-agent systems with event-triggered cloud access. In: European control conference, pp. 954–961 (2015)
7. Nowzari, C., Pappas, G.: Multi-agent coordination with asynchronous cloud access. In: American Control Conference, pp. 4649–4654 (2016)

8. Lei, Z., Zidong, W., Donghua, Z.: Event-based control and filtering of networked systems: a survey. *Int. J. Autom. Comput.* **14**(3), 239–253 (2017)
9. Abdelaal, A. E., Hegazy, T., Hefeeda, M.: Event-based control as a cloud service. In: American Control Conference IEEE, pp. 1017–1023 (2017)
10. Nankun, M., Xiaofeng, L., Tingwen, H.: Event-based consensus control for a linear directed multiagent system with time delay. *IEEE Trans. Circ. Syst. II Express Briefs* **62**(3), 281–285 (2017)
11. Jian, Y., Hongjin, Z., Hongjun, C., Weidong, Z.: Output event triggered consensus control of nonlinear multi-agent systems with relative state constraints. *ISA Trans.* **108**(3) (2020)
12. Aranda-Escalastico, E., Rodriguez, C., Guinaldo, M., et al.: Asynchronous periodic event-triggered control with dynamical controllers. *J. Franklin Inst.* **355**(8), 3455–3469 (2018)
13. Xiangyu, M., Lihua, X., Yeng-Chai, S.: Asynchronous periodic even-triggered consensus for multi-agent systems. *J. Automatica.* **84**, 214–220 (2017)
14. Jinjie, H., Xiaozhen, P., Xianzhi, H.: Event-triggered and self-triggered  $H\infty$  output tracking control for discrete-time linear parameter-varying systems with network-induced delays. *Int. J. Syst. Sci.* 1–20 (2020)
15. Wenfeng, H., Lu, L., Gang, F.: output consensus of heterogeneous linear multi-agent systems by distributed event-triggered/self-triggered strategy. *IEEE Trans. Cybern.* **47**(99), 1914–1924 (2017)
16. Almeida, J., Silvestre, C., Pascoal, A.: Synchronization of multiagent systems using event-triggered and self-triggered broadcasts. *IEEE Trans. Autom. Control* **62**(9), 4741–4746 (2017)
17. Zeyu, H., Wallace-K, T., Qiang, J.: Event-triggered synchronization for nonlinear multi-agent systems with sampled data. circuits and systems I: regular papers. *IEEE Trans.* (99), 1–9 (2020)
18. Dimarogonas, D., Frazzoli, E.: Distributed event-triggered control for multi-agent systems. *IEEE Trans. Autom. Control* **457**, 1291–1297 (2012)
19. Yongfeng, G., Liu, L.: Lyapunov-based triggering mechanisms for event-triggered control. *Int. J. Control Autom. Syst.* **18**(5) (2019)
20. Bo, Z., Xiaofeng, L., Tingwen, H., Guo, C.: Pinning exponential synchronization of complex networks via event-triggered communication with combinational measurements. *Neurocomputing* **157**, 199–207 (2015)

# Towards Distractibility Induced trust Management Using BlockChain for Edge Computing



Haochen Yang , Guanghui Wang , Lifeng Dong , and Xin He

**Abstract** Edge computing solves massive data processing problems by utilizing the storage and computing functions of edge devices. Due to the openness of edge computing, distracted devices inevitably participate in edge computing and affect the network computing performance, which degrades trust management of edge devices. To solve the above issue, this paper proposes a trust management framework based on blockchain for edge computing scenarios (ECTMF). First, the ECTMF framework slices users and provides corresponding services based on different requirements. It evaluates the trustworthiness of end devices and distinguishes distracted devices in edge computing. Then, the Proof of Monitor (PoM) consensus mechanism is proposed to continuously monitor distracted end devices in edge computing. Simulations validate the effectiveness and efficiency of the ECTMF and PoM.

**Keywords** Edge computing · Trust management · BlockChain · Consensus mechanism

## 1 Introduction

With the increasing number of portable smart devices, huge amounts of data need to be processed. As a promising mode of computing, edge computing can solve massive

---

H. Yang · G. Wang · X. He (✉)  
School of Software, Henan University, Kaifeng, China  
e-mail: [hexin@henu.edu.cn](mailto:hexin@henu.edu.cn)

H. Yang  
Henan Provincial Engineering Research Center of Intelligent Data Processing, Kaifeng, China

G. Wang  
Henan International Joint Laboratory of Intelligent Network Theory and Key Technology, Henan University, Kaifeng, China

X. He  
Institute of Intelligent Network System, Henan University, Kaifeng, China

L. Dong  
Henan Jiuyu Tenglong Information Engineering Co., Ltd., Zhengzhou, China

data processing by utilizing the storage and computing functions of edge devices [1]. Edge computing usually divides users into different types by network slicing, and provides corresponding services according to different needs [2].

Openness is a typical characteristic of edge computing [3], which inevitably introduces distracted end devices. The end devices in edge computing are defined as distracted devices when it performs tasks at a rate lower than its computing capacity, termed as *device distractibility*. The device distractibility decreases the quality of the edge computing tasks. For example, the network computing performance is decreased because the distracted end devices execute edge computing tasks negatively. Trust management is an effective way to evaluate the trustworthiness of end devices in edge computing, which can be used to detect distracted end devices. Therefore, it is an important problem to research on how to manage the trust of the distracted end devices so as to accurately identify distracted end devices in edge computing.

Traditional trust management is centralized. The information of users in centralized trust management is usually stored in a central server [4]. The central server existence a single point of failure problem. The process of centralized trust management is not immutable and traceable. A center point will tamper information of users for their interests. It is a desire to design a decentralized trust management framework in edge computing.

Blockchain is an effective technology for decentralized trust management [6]. The blockchain technology has been successfully applied to the bit-coin system [5]. In the blockchain, the transaction information is stored in blocks, and the hash point is used to link every block [7]. The blockchains is composed of blocks. All users jointly maintain the blockchain. Decentralization, immutability, and traceability are features of blockchain. At the same, the user publishing blocks is called miner. The system selects the miner's block by the consensus mechanism.

In this paper, the blockchain is utilized as a ledger of trust values to achieve decentralized trust management, where the trustworthiness is calculated using the concept of device distractibility. A consensus mechanism is proposed based on the number of distracting behaviors, where the block structure is designed for the trust value blockchain. Therefore, a blockchain based trust management framework is presented to evaluate the trustworthiness of end devices in edge computing, which is called Edge Computing Trust Management Framework (ECTMF). A novel consensus mechanism Proof of Monitor (PoM) is designed based on an alliance chain to monitor end devices that negatively execute edge computing tasks. The contributions are summarized as follows:

- (1) We propose the ECTMF framework to manage and evaluate the distractibility induced trust of the end devices in edge computing. The ECTMF framework can identify the distracted end devices effectively so as to improve the efficiency of performing the edge computing tasks.
- (2) We propose a novel consensus mechanism PoM for the ECTMF framework, which can monitor distracted end devices continuously. The PoM is designed for alliance chain. The PoM can monitor end devices effectively and possessing no power consumption.

- (3) We design the block structure for the ECTMF framework by adding the number of distracting behaviors in the block header. Simulation experiments are carried out to demonstrate the superiority of performance for the proposed ECTM and PoM.

The rest of this paper is arranged as follows. Section II is the related work. Section III proposes the ECTMF framework and the PoM mechanism. Performance evaluation is given in section IV and section V is the conclusion of this paper.

## 2 Related Work

This chapter analyzes the related work of trust management and blockchain. This paper analyzes related work and designs a trust management framework based on blockchain for edge computing scenarios. This blockchain-powered framework evaluates the trustworthiness of end devices and distinguishes distracted devices in edge computing.

### 2.1 *Trust Management*

Trust management has been researched for many years. Trustworthiness is defined as a firm belief in the ability of an entity to act reliably, safely, and reliably in a specific environment in [8]. Yang et al. in [7] proposed a trust management scheme in vehicular networks. They use the Bayesian inference formula to calculate the offset of vehicle trust value. In [9], Feng et al. overcome the attacks in trust evaluation by improving the feedback mechanism. [10] proposed a decentralized framework for crowdsourcing based on blockchain. However, the framework lacks a trust management approach to evaluate the trustworthiness in the network. In edge computing, the distracted end devices affect the computing performance of the network. Thus, trust management in edge computing is essential to distinguish the distracter end devices. By managing trust in distracted devices, edge computing can effectively improve the efficiency of task execution.

### 2.2 *Blockchain*

Blockchain was first proposed in [5]. Blockchain has been applied in various scenarios, such as vehicle networks [7], Internet of Things (IoT) [12], etc. In [11], An et al. consider privacy in node selection based on twice consensuses of blockchain. Boussard et al. in [13] proposed a reputation management system for IoT, they evaluate the reliability of each router and avoid the shortcoming of the centralized

system by blockchain. [7] aims at the problem of slow update speed of effective information in blockchain proposed a novel consensuses mechanism. This consensuses mechanism is a joint of PoW and PoS, which makes effective information update more timely. This paper proposes a novel consensuses mechanism PoM, which can monitor distracted end devices. However, the existing research on consensus mechanisms for blockchains cannot continuously monitor distracted end devices in edge computing. To improve the efficiency of edge computing, a consensus mechanism that can effectively distinguish distracted devices in edge computing is needed.

### 3 System Model and Problem Setup

#### 3.1 System Model

The ECTMF focus on evaluating and managing trustworthiness of end devices in edge computing. The participants in the network are sliced. Intelligent devices and mobile terminals in edge computing are sliced into end devices. The end devices are used to execute edge computing tasks. The  $i$ th edge computing device is expressed as  $D_i$  in this paper. The edge computing tasks contain data storage, data collection, data processing, etc. The  $i$ th edge computing tasks in the network is expressed as  $T_i$ . The change of  $D_i$ 's trustworthiness in  $T_k$  is expressed as  $offset_i^k$ . Devices with great computing power (i.e., edge computing devices) such as base stations are sliced into miners. Miners mainly fulfill to send and receive task information. At the same time, miners jointly maintain a blockchain, which stores trustworthiness information of end devices in edge computing. The  $i$ th miner in the network is expressed as  $M_i$ .

#### 3.2 Problem Setup

This paper proposes a decentralized trust management framework based on blockchain. The goal of this framework is to evaluate and manage the trustworthiness of edge computing. Moreover, a novel consensuses mechanism is proposed to monitor distracted end devices. The design goals mainly include:

- (1) Accuracy. The ECTMF can evaluate the trustworthiness of edge computing end devices accurately by effective information in blockchain. Simulation experiment verifies that the evaluated trustworthiness is accurate and fast.
- (2) Timeliness. This paper proposes a novel consensuses mechanism PoM, which can monitor distracted end devices continuously. The block contains information on distracted end devices that can be more quickly publish in the blockchain.
- (3) Decentralization. The trust management framework is decentralized in this paper. ECTMF uses blockchain to record the trustworthiness of end devices

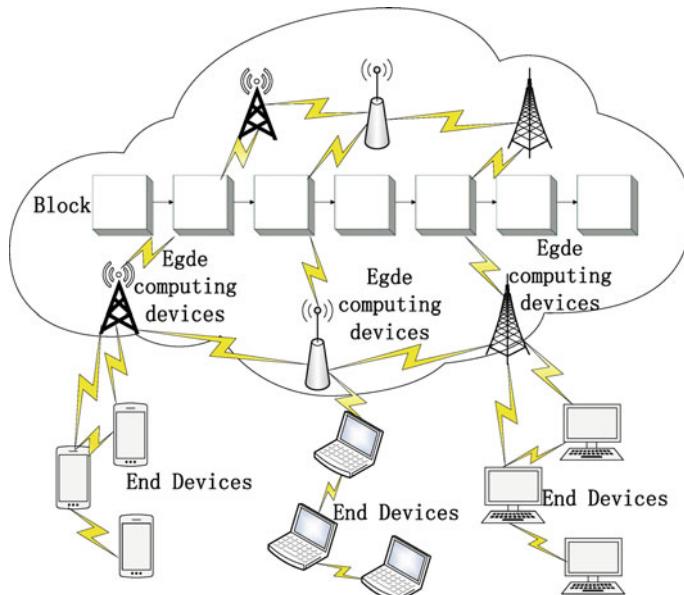
in edge computing. Blockchain is a distributed ledger, which not needs a trusted third party. Based on blockchain, the evaluation of trustworthiness in the ECTMF is decentralized.

## 4 Edge Computing Trust Management Framework

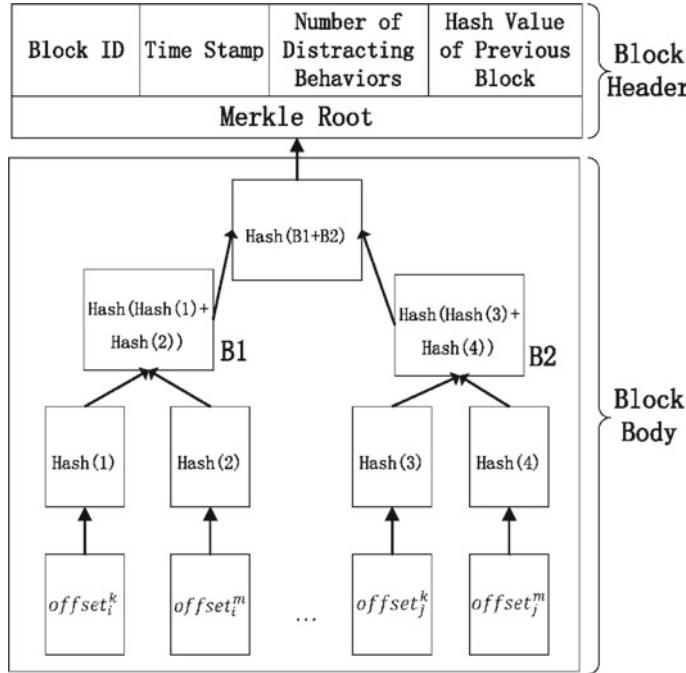
### 4.1 ECTMF

ECTMF provides a trustworthiness calculation method, which evaluates end devices' concentration degree. Then a novel consensus mechanism PoM is proposed to monitor distracted end devices continuously. The ECTMF is shown in Fig. 1. The ECTMF composes of end devices and miners. In Fig. 1, miners are base stations. They record trustworthiness information of end devices and jointly maintain a blockchain that is used for storing trustworthiness information. Miners continuously monitor the edge computing task information in the network. They take the trustworthiness information of end devices into a block and broadcasts it. Miners who receive the block will continue to monitor for  $t$  seconds. After monitor, the miner sorts  $m$  blocks he received in  $t$  seconds. Then add the top-ranked block to the end of the blockchain he maintains. The end devices accept and complete edge computing tasks.

The structure of the block in ECTMF is shown in Fig. 2. In ECTMF, a block contains block header and block body. Block header contains block id, time-stamp,



**Fig. 1** Edge computing trust management framework



**Fig. 2** Structure of block

number of distracting behaviors, hash the value of previous block. Where the number of distracting behaviors is used in the PoM, which is described in the consensuses mechanism part. The trustworthiness information of end devices is stored in the block body in the form of the Merkle Tree. The root of the Merkle Tree is called the Merkle Root.

## 4.2 Trustworthiness Evaluation in Edge Computing

Trustworthiness in this paper is end devices' concentration degree when executing edge computing tasks. Distracted end devices complete tasks at a speed slower than their computing power. In ECTMF, end devices get a trust value offset in every edge computing task. The data of trust value offset is stored in blockchain, and end devices can get device trustworthiness by cumulating all their trust value offset. Suppose the real computing power of devices can be obtained in advance. The real computing power is the ability to perform a particular task. This hypothesis is valid because the hardware attributes of devices are open source. Based on this hypothesis,  $D_i$ 's trust value offset in  $T_k$  is expressed as Formula 1. The parameter  $\alpha$  is expressed as Formula 2.

$$offset_i^k = \alpha(P_i^k - P_i^{exp}) / P_i^{exp} \quad (1)$$

$$\alpha = 1/(num_i^{task})^2 \quad (2)$$

where  $P_i^k$  is computing power that  $D_i$  used to execute  $T^k$ .  $P_i^{exp}$  is true computing power of  $D_i$ .  $(P_i^k - P_i^{exp})/P_i^{exp}$  called concentration degree of  $D_i$ .  $\alpha$  is a parameter to adjust the weight of trust value offset. If  $D_i$ 's concentration is certain, with the number of tasks increases, the trustworthiness of  $D_i$  tends to be stable. Therefore, the value of  $\alpha$  decrease with the number of task increases.

Data of  $offset_i^k$  is stored in the blockchain, other users get  $D_i$ 's trustworthiness by accumulating  $offset_i^x$ . Where  $x$  is  $D_i$  completed tasks. The trustworthiness of  $D_i$  in range  $[0, 1]$ . Let  $Tru_i$  denote the trustworthiness of  $D_i$ ,  $Tru_i$  is expressed as Formula 4.

$$Tru_i = \min \left[ \max \left[ \left( 1 - \sum_{k=1}^n offset_i^k \right), 0 \right], 1 \right] \quad (3)$$

### 4.3 Proof of Monitor

Miners pack trustworthiness information of end devices (i.e., trust value offset of devices) into a block and broadcast it. Miners who receive the block will verify the block. If the block is verified to be qualified, the miner will continue to listen for  $t$  seconds. Suppose this miner receives  $b - 1$  blocks in  $t$  seconds, then sort  $b$  blocks by Algorithm 1. Finally, the miner adds the block with the highest priority to the end of the blockchain. Algorithm 1 describes the proof of monitor.

---

**Algorithm 1 Proof of Monitor**

---

**Require:**  $b$  valid blocks( $block_1; block_2...block_b$ )

**Ensure:** 1 block with the highest priority after sorting

```

 $DB_i = \max(DB_1, DB_2, \dots, DB_b)$ 
if  $DB_i \neq (DB_1, DB_2, \dots, DB_{i-1}, DB_{i+1}, DB_b)$  then
    out put  $block_i$ 
end if
if  $DB_i = DB_j = \dots = DB_k$  then
    Compare the number of blocks published by miners who generated  $block_i$ ,
     $block_j, \dots, block_k$ . Output the block packed by the miner with the least
    number of published blocks.
else
    Output  $block_i$  with the largest hash value.
end if
```

---

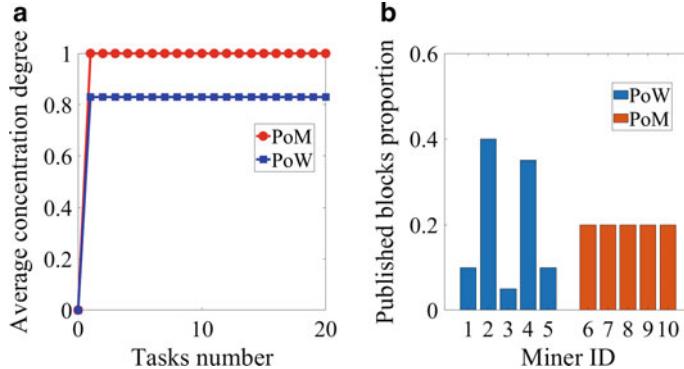
In proof of monitor, miners sort  $b$  blocks by the number of distracting behaviors recorded in blocks. Distracting behaviors are defined as end devices perform tasks rate lower than its' computing power, i.e.,  $P_i^k - P_i^{exp} < 0$ . The number of distracting behaviors recorded in block  $i$  is expressed as  $DB_i$ . Miners compare the number of distracting behaviors first. Output block with the most number of distracting behaviors. If the number of distracting behaviors in  $b$  blocks is identical, output the block packed by the miner who published blocks least. It can reduce centralization in the blockchain. If these steps cannot output a unique block. Algorithm 1 compares the hash value of these blocks and output the block with the most hash value. PoM can monitor distracted end devices in edge computing, so as to improve the efficiency of edge computing. Compared with PoW, PoM has no power consumption.

## 5 Performance Evaluation

This experiment is implemented on an HP notebook with an Intel Core i7-8750 h processor and 8.00 Gb (7.88 Gb available) memory. The number of miners, end devices, and tasks is set 5, 100 and 20, respectively.

### 5.1 Evaluation Results of ECTMF and PoM

Simulation experiments are conducted to evaluate the performance of ECTMF and PoM. Set the concentration degree of 50 end devices are 0.5 and others are 1.0. The trustworthiness of 100 end devices is updated by PoW and PoM under ECTMF. Select 50 end devices by the greedy algorithm, the average concentration degree of selected



**Fig. 3** Evaluation results of ECTMF and PoM

end devices is shown in Fig. 3a. At the same time, the simulation experiment evaluates the centralization degree of PoM. Set 5 miners to occupy all computing power, and the computing power percentage of miners is random (i.e., [0.1,0.4,0.05,0.35,0.1]). All end devices' concentration degrees are set 1.0. Miners ID 1–5 use the PoW, and miners ID 6–10 use the PoM. After several consensuses, published blocks proportion of 10 miners is shown in Fig. 3b.

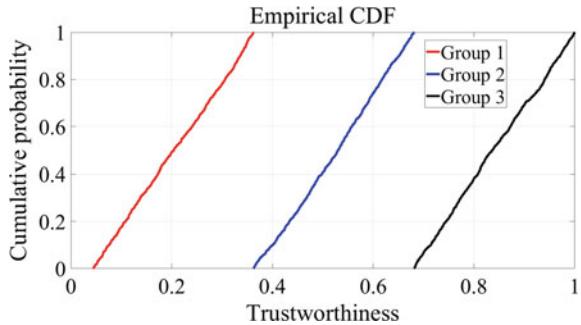
In Fig. 3a, the average concentration of 50 end devices under PoM is 1.0, and under PoW is 0.83. It can be seen, the PoM can discover distracted end devices and increase the average concentration degree of edge computing timely. Without the ECTMF, the result is equal PoW. At the same time, ECTMF as a decentralized framework can avoid centralized server problems. When all end devices are concentrated, the comparison of PoW and PoM is shown in Fig. 3b. After several consensuses, the proportion of the published blocks of PoW is dependent on the computing power of miners. The proportion is equal under PoM. It can be seen that computing power does not affect on the number of published blocks under PoM. PoM can reduce centralization in the blockchain.

## 5.2 Evaluation Results of Trustworthiness Evaluate

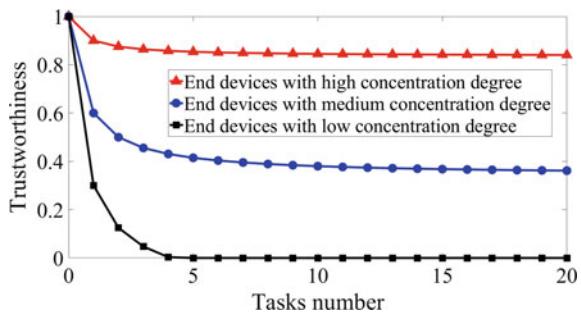
This paper uses ECTMF to evaluate the trustworthiness of three group end devices. Sets 1000 end devices for every group. The concentration degree of the first group is set (0.4, 0.6]. The second group is set (0.6, 0.8]. The third group is set (0.8, 1]. After 20 tasks, the trustworthiness cumulative probability of 3000 end devices is shown in Fig. 4.

It can be seen ECTMF can evaluate trustworthiness exactly. The trustworthiness can accurately reflect concentration degree. The trustworthiness of the first group device is less than the second group and far less than the third group.

**Fig. 4** Trustworthiness cumulative probability of three groups devices



**Fig. 5** Evaluation results of ECTMF trust management



Trustworthiness evaluation of three kinds of end devices (i.e., devices with high concentration degree, medium concentration degree and, low concentration degree) is carried out. Set the concentration degree of the device with a high concentration degree is 0.9, the device with a medium concentration degree is 0.6 and the device with a low concentration degree is 0.3. The trustworthiness change of three kinds of end devices in 20 tasks is shown in Fig. 5.

In Fig. 5, it can be seen that ECTMF can quickly evaluate the trustworthiness of end devices in edge computing. The trustworthiness tends to be stable as the number of tasks increases when the concentration degree of the device is changeless. After 20 tasks, the trustworthiness of end devices with high concentration degree is higher than medium concentration degree devices. The trustworthiness of end devices with low concentration degree tends to 0.

## 6 Conclusion

This paper proposes a distractibility induced trust management framework ECTMF for edge computing, which is able to evaluate the effect of the distractibility on the trustworthiness. A consensus mechanism proof of monitor PoM is presented for

blockchain which stores trustworthiness information for ECTMF. The PoM mechanism does not need power consumption and can monitor distractibility continuously. The network computing efficiency is improved by using PoM. The simulation experiment verifies that ECTMF can improve the average concentration degree of end devices and reduce the centralization of blockchain.

**Acknowledgements** This work was supported by the Elite Postgraduate Students Program of Henan University (SYL20060174), the Key Technologies Research and Development Program of Henan under Grant 212102210078, 212102210090, and 212102210094, the Henan Provincial Major Public Welfare Projects under Grant 201300210400, the China Postdoctoral Science Foundation under Grant 2020M672211, and the Key Scientific Research Projects of Henan Provincial Colleges and Universities under Grant 21A520003.

## References

1. Hnab, C., et al.: Heterogeneous edge computing open platforms and tools for internet of things. *Futur. Gener. Comput. Syst.* **106**, 67–76 (2020)
2. Stefan, R., et al.: Network slicing. *5G Core Networks* (2020)
3. Fuhong, L., et al.: Security issues in emerging edge computing. *China Commun.* **17**(10), 4–5 (2020)
4. Haiying, M., et al.: Blockchain-based mechanism for fine-grained authorization in data crowdsourcing. *Futur. Gener. Comput. Syst.* **106**, 121–134 (2020)
5. Satoshi Nakamoto Institution, Bitcoin: A Peer-to-Peer Electronic Cash System. <http://www.bitcoin.org/bitcoin.pdf>. Last accessed 3 Nov 2008
6. Rajat, C., et al.: BEST: blockchain-based secure energy trading in sdn-enabled intelligent transportation system. In: *Computers & Security*, vol. 85, pp. 288–299 (2019)
7. Zhe, Y., et al.: Blockchain-based decentralized trust management in vehicular networks. *IEEE Internet Things J.* 1–1 (2018)
8. Grandison, T., Sloman, M.: A survey of trust in internet applications. *IEEE Commun. Surv. Tutorials* **3**(4), 2–16 (2000)
9. Feng, W., Yan, Z.: MCS-Chain: Decentralized and trustworthy mobile crowdsourcing based on blockchain. *Futur. Gener. Comput. Syst.* **95**, 649–666 (2019)
10. Ming, L., et al.: CrowdBC: a blockchain-based decentralized framework for crowdsourcing. *IEEE Trans. Parallel Distrib. Syst.* 1–1 (2018)
11. Jian, A., et al.: TCNS: node selection with privacy protection in crowdsensing based on twice consensuses of blockchain. *Network and Service Management. IEEE Trans. Netw. Serv. Manag.* **16**(3), 1255–1267 (2019)
12. Min, L., Helen, T., Xianbin, W.: Mitigating routing misbehavior using Blockchain-based distributed reputation management system for IoT networks. In: *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6. IEEE (2019)
13. Mathieu, B., Serge, P., Pierre, P., Matteo, S., Erez, W.: STewARD: SDN and blockchain-based trust evaluation for Automated Risk management on IoT Devices. In: *38th IEEE International Conference on Computer Communications Paris*, pp. 841–846 (2019)

# Social Robot Navigation Based on a 2D Gauss-Gumbel Spatial Density Model in Human-Populated Environments



Jianfang Lian , Wentao Yu , Kui Xiao , Feng Qu , and Chaofan Liu

**Abstract** Social robot navigation must consider not only task constraints, such as the minimum path length, but also social conventions, such as satisfying the social acceptability of the path. This paper presents a new strategy for social robot navigation based on 2D Gauss-Gumbel spatial density function to consider the human state (position, direction and motion) and social interaction information related to robots, which model the personal space and social interaction space respectively. The personal space and social interaction space constitute a Dynamic Social Space (DSS). The DSS based human comfort and safety navigation can estimate the approaching pose of a robot for a person or a group of people, so the robot can ensure not only people's safety but also comfort when approaching a person or group of people in social situations. We evaluate the developed model through simulation and real-world experiments using the newly proposed social individual comfort index and social group comfort index.

**Keywords** 2D Gauss-Gumbel spatial density function · Socially aware robot navigation · Dynamic social space · Individual and group comfort indices

## 1 Introduction

The pursuit of reducing labor and a more comfortable life is the development goal of human society [1]. Mobile robots have gradually expanded from factories and production workshops with a single environment to complex and changeable indoor environments [2]. Navigation is the prerequisite for the autonomy of social robots. Therefore, the ability of social robots to adjust their behavior according to social convention is the key to social robot navigation [3].

Prototype engineering dates back at least 20 years ago, when human safety was a primary consideration. Traditional navigation technology usually treats humans as

---

J. Lian · W. Yu (✉) · K. Xiao · F. Qu · C. Liu

Central South University of Forestry and Technology, Changsha 410004, Hunan, China

e-mail: [wtyu\\_csuft@126.com](mailto:wtyu_csuft@126.com)

ordinary obstacles. Commonly used navigation techniques include proactive avoidance [4], A \* algorithm [5], Artificial Potential Field Algorithm (APF) [6] and Bidirectional Random Extended Tree Algorithm (BRRT) [6]. However, these methods do not satisfy human social acceptability.

The concept of personal space is a widely studied aspect of the social customs of spatial interaction [7]. The main idea is to create a comfortable space acceptable to humans during robot navigation. Therefore, in recent years, the number of works incorporating this personal space model concept in mobile robot navigation has increased [8].

Papadakis et al. proposed that humans model the human's personal space as a probability function of nonlinear proportions in the robot state space, and design the structure and shape of the interactive space through the kernel principal component analysis (KPCA) algorithm [9]. In this way, complex social interactions of any shape and size can be realistically simulated. Thereby predicting the approaching trajectory of people. However, this model only applies to a pair of interacting people. And the model does not consider the person's direction, that is to say, models built by people in different directions are the same.

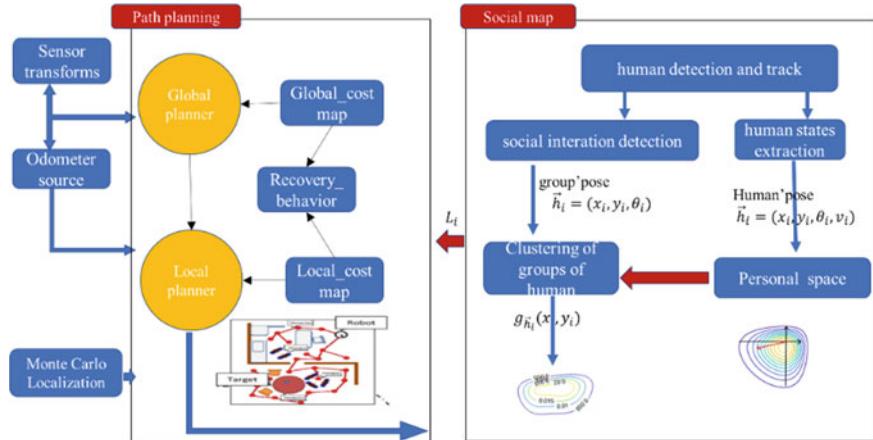
In this paper, the human interaction space is modeled with a nonlinear scale probability function. The 2D Gauss-Gumbel spatial density function was used to describe the extended personal space and social interaction space according to the individual state (location, direction, and velocity) of personal information and social group information, respectively. The personal space and social interaction space constitute a Dynamic Social Space (DSS), which considers the safety and psychological comfort of humans.

The DSS based human comfort and safety navigation can estimate the approaching pose of a robot for a person or a group of people, so the robot can ensure people's also comfort. On the basis of the comfort distance of waypoints proposed by Hall, this article proposes two intuitive metrics to measure human safety and comfort—Personal Comfort Index (PCI) and Group Comfort Index (GCI). The validity of the structure of social interaction space is verified.

## 2 System Overview

Navigation is the prerequisite for mobile robot autonomy. The goal of social robot navigation is to make the behavior of mobile robots meet certain social conventions, thereby improving the social acceptability of planned paths. In order to achieve this goal, this paper developed a navigation framework for social robots, as shown in Fig. 1.

In the first part, the social map aims to distinguish humans from conventional obstacles by extracting the social space-time features near the robots to develop a dynamic social space map and predict the robot's approach to humans or a group of people.



**Fig. 1** Navigation framework for social robots

The second part is path planning, which refers to planning a collision-free optimal path based on the surrounding environment information through a path planning algorithm to reduce the energy consumption of the robot. This paper combines global path planning and local path planning to generate paths. Global path planning uses A\* algorithm, local path planning uses DWA algorithm.

The third part is mobile robot positioning. We derive the Monte Carlo localization algorithm.

Finally, this paper proposed two indicators to measure human comfort: Personal Comfort Index (PCI) and Group Comfort Index (GCI).

### 3 Social Map

The social map proposed in this article is a bottom-up approach to solve this problem. The traditional approach is to solve this problem through a top-down perspective. From a top-down perspective, both humans and robots are regarded as one point, which means that the distance between robots and humans and between humans is constant. This method of treating both robots and humans as mass points does not satisfy humans' posture constraints. The dynamic social interaction space can further promote the recognition and reasoning of mobile robots. Social maps have comprehensively improved the situational awareness of robots. The construction of a dynamic social interaction space needs to meet the following three goals,

1. Accurately describe the human psychological comfort space.
2. Grasp the dynamic changes of the state of social space
3. Simulated crowd social interaction area.

### 3.1 Human Model

We describe the robot operating environment as a shared state space between humans and robots, and humans are a vector in this space. The human body state obtained from sensor information and visual information includes position, speed, and orientation. Let  $S \in R^2$  be the space of the Global Map. An individual  $i$  is represented by its pose (position, orientation, and velocity),  $h_i = [x_i, y_i, \theta_i, v_i]^T$ , being  $[x_i, y_i]^T \in S$ , and  $\theta_i \in [0, 2\pi]$ .

### 3.2 Personal Space Model

This paper uses  $h_i$  as the underlying characteristic input to model the personal space, and selects a single functional function with the following ideal attributes to construct the personal space:

- Smoothness. Smoothness is designed to allow the robot to respond smoothly to changes in human social sensitivity.
- Velocity dependent. The study of human–human and human–robot interactions has inspired a reliance on human speed.
- Directionally dependent. Similarly, the study of human–human and human–robot interactions has inspired a reliance on human direction.
- Sealing ability. Closure means the integrity of personal space. Closure is mainly to ensure the absolute safety of human in the process of robot movement.

The one-dimensional Gaussian probability density function is symmetric and does not satisfy the directional dependence of personal space. The Gumbel probability density function is monotonous and does not satisfy the closed character of personal space. Therefore, we proposed two-dimensional Gauss-Gumbel probability density function, which is composed of a two-dimensional Gumbel probability density function and a one-dimensional Gaussian probability density function.

$$g(X, Y) = \frac{G(X, Y)}{\sigma_1 \sigma_2 \sigma_3 \sigma_4} * A^{\frac{1}{a}} * B^{\frac{1}{a}} * \left( A^{\frac{1}{a}} + B^{\frac{1}{a}} \right)^{(a-2)} * \left[ a * \left( A^{\frac{1}{a}} + B^{\frac{1}{a}} \right) - (a-1) \right] \quad (1)$$

$$G(X, Y) = e^{-\left( e^{-\frac{X-\bar{x}}{\sigma_3}} + e^{-\frac{Y-\bar{y}}{\sigma_4}} \right)^a} \quad (2)$$

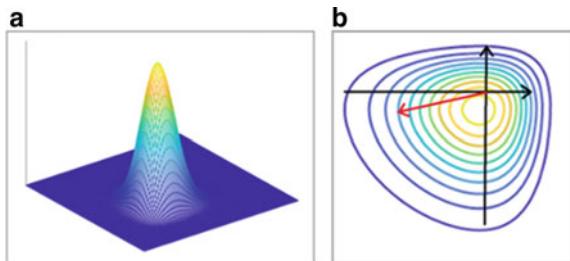
where  $A = e^{-\frac{(X-\bar{x})^2}{\sigma_1^2}}$ ,  $B = e^{-\frac{(Y-\bar{y})^2}{\sigma_2^2}}$ ,  $\sigma_1 = \sigma_2 = 1$ ,  $a = 1$ . The size of  $\sigma_3$  and  $\sigma_4$  determines the shape of personal space.

When a person is stationary,  $\sigma_3 = \sigma_4 = 1$ .

When a person is in motion, the average human walking speed is 1.25 m/s. Therefore,  $v = 1.25$  m/s,  $\theta \in (0, 360^\circ)$ .

If  $0 \leq \theta < \pi$ ,

**Fig. 2** The contour and surface maps of a 2D Gumbel-Gauss spatial density function



$$\sigma_3 = v * \sqrt{|\sin(\theta)|} * \sqrt{|\sin(\theta)|} * \sqrt{|\sin(\theta)|} \quad (3)$$

$$\sigma_4 = v * \sqrt{|\cos(\theta)|} * \sqrt{|\cos(\theta)|} * \sqrt{|\cos(\theta)|} \quad (4)$$

If  $\pi \leq \theta < 3\pi/2$

$$\sigma_3 = v * \sin(\theta) * \sin(\theta) * \sin(\theta) \quad (5)$$

$$\sigma_4 = v * \cos(\theta) * \cos(\theta) * \cos(\theta) \quad (6)$$

If  $3\pi/2 \leq \theta < 2\pi$

$$\sigma_3 = v * \cos(\theta) * \cos(\theta) * \cos(\theta) \quad (7)$$

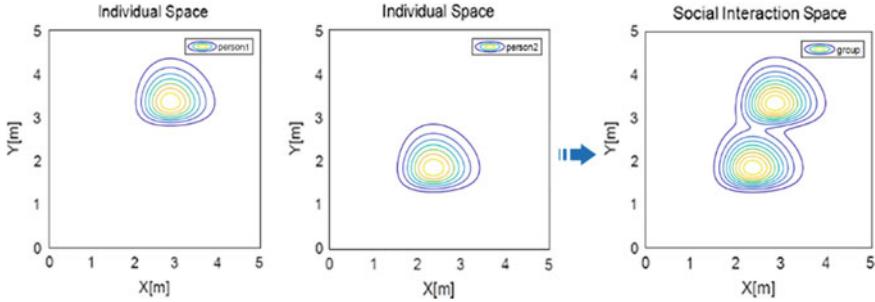
$$\sigma_4 = v * \sin(\theta) * \sin(\theta) * \sin(\theta) \quad (8)$$

Figures 2a, b show the contour and surface maps of a 2D Gumbel-Gauss Spatial Density Function, respectively.

The red arrow indicates the direction of an individual and the outermost blue circle represents personal space of an individual.

### 3.3 Social Interaction Space

There are many small groups in the social environment, that is, interacting with others. The number of members in a small group is generally between 2 and 4 people. In traditional social practice, robots are not allowed to pass through in the interactive group. This paper uses sensor information and visual information to obtain group information around the robot. Based on the 2D Gauss-Gumbel social probability density function to model the individual space, this paper uses the mathematical



**Fig. 3** Construction process of social interaction space

average value to set the position of the group, and makes the mathematical combination of the individual space in the group, so as to complete the modeling of the social space of the group.

Thus, a human group social interaction space is defined as follows. Let  $g_{h_i}(x_i, y_i)$  be the personal space function for each individual  $h_i$  in the set of all  $P$  of all people in S. a human group social interaction space  $G(P)$  is defined as follows.

$$G(P) = \text{contour} \left( \frac{X_1 + X_2 + \dots + X_i}{i}, \frac{Y_1 + Y_2 + \dots + Y_i}{i}, \frac{[g(X_1, Y_1) + g(X_2, Y_2) + \dots + g(X_i, Y_i)]}{i} \right) \quad (9)$$

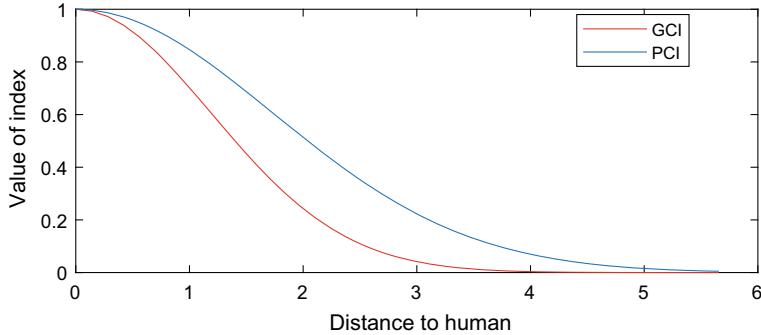
Contour is a contour function, and the construction process of social interactive space is shown in Fig. 3.

### 3.4 Evaluation Index

To validate the proposed model, we proposed two metrics to measure socially acceptable behaviour soft the mobile robot: Personal comfort index (PCI), Group comfort index (GCI). The PCI value is used to measure the comfort of individual, while GCI value is used to measure group psychological comfort.

$$\text{PCI} = \max_{i=1:N} \exp \left( - \left( \left( \frac{x_r - x_i^p}{\sigma_0^{px}} \right)^2 + \left( \frac{y_r - y_i^p}{\sigma_0^{py}} \right)^2 \right) \right) \quad (10)$$

$(x_r, y_r)$  is the path point of the mobile robot;  $(x_i^p, y_i^p)$  is the position of the  $i$ th person on the global map S, N is the number of people in the global map.  $\sigma_0^{py} = \sigma_0^{px} = 5$



**Fig. 4** The variation curve of rating value with distance from others

$$GCI = \exp\left(-\left(\left(\frac{x_r - x_o^p}{\sigma_k^{px}}\right)^2 + \left(\frac{y_r - y_o^p}{\sigma_k^{py}}\right)^2\right)\right) \quad (11)$$

$(x_o^p, y_o^p)$  is the center position of the  $o$  crowd in the global map S; M is the number of people in the global map  $\sigma_k^{py} = \sigma_k^{px} = 5$ . The change curve of evaluation index and distance to people is shown in Fig. 4.

The PCI value ranges from 0.0 to 1.0, where the closer the distance between the mobile robot and the human, the higher the PCI value. When the relative distance between the human and the robot is greater than 6 m, the PCI value is approximately equal to zero. The smaller the PCI value, the more comfortable humans feel. According to the comfortable distance proposed by Hall, the PCI threshold  $T_c = 0.9$  is set.

Similar to PCI, GCI values range from 0.0 to 1.0. As shown in Fig. 4, when the robot moves farther from the center of the social space, the value of GCI decreases. The shorter the distance between the robot and the center of the social group, the higher the GCI value, indicating that the robot's behavior is not accepted by the society in the social environment. According to the comfortable distance proposed by Hall, the GCI threshold  $T_p = 0.7$  is set.

## 4 Experiment

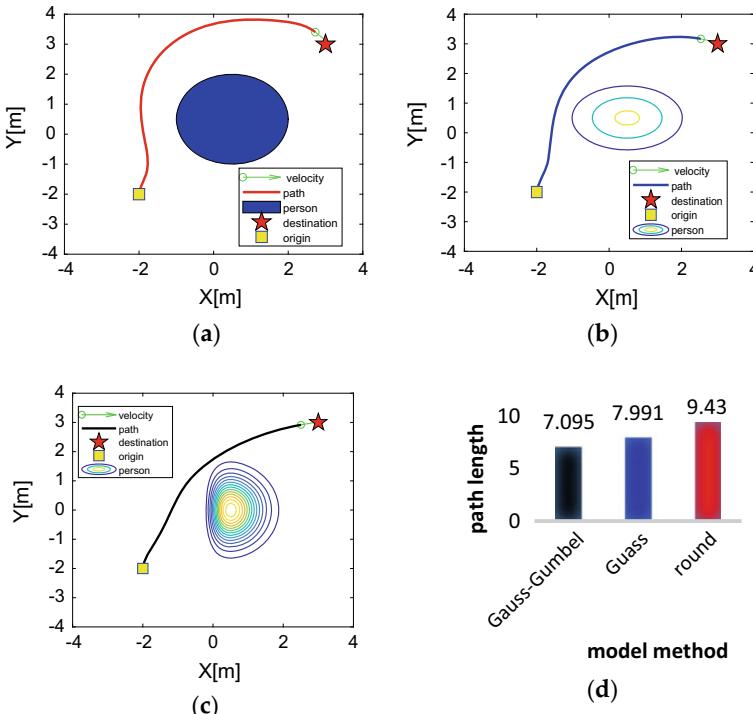
In this section, we model different kinds of groups of people, including two persons walked towards each other, two walked in company, one chased one, and three stood talking.

## 4.1 Personal Space

In this section, we conducted three sets of experiments. The first set of experiments modeled humans as ordinary circular obstacles, the second set of experiments modeled humans as Gaussian probability density functions, and the third set of experiments modeled humans as Gauss-Gumbel spatial density function. All experimental path planning algorithms use the A\*-DWA algorithm, and the starting point, target point and initial state of the robot are consistent. In addition, the information of people in all experiments is the same.

Figure 5a shows results of treating humans as ordinary circular obstacles, Fig. 5b shows the results of modeling humans with 2D Gauss-Gumbel spatial density function, and Fig. 5c shows the results of modeling humans with Gaussian-Gamber social probability density function. Experimental results. Figure 5d shows the path analysis results of the three sets of experiments.

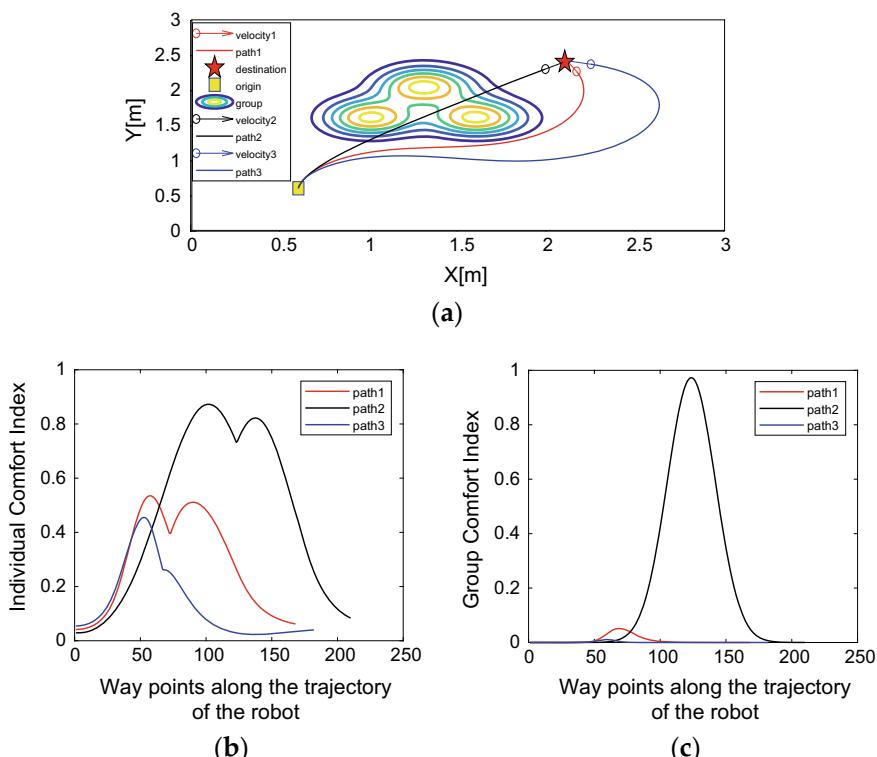
As shown in Fig. 5, the model method we proposed saves working space on the basis of meeting social conventions, the planned path length is the shortest, and the efficiency of navigation is improved.



**Fig. 5** The variation curve of rating value with distance from others

## 4.2 Social Interaction Space

In this section, simulation experiments are a group of three standing people. To test the feasibility of the model constructed in this paper, we also tested the pros and cons of the DWA algorithm and the fusion DWA-A\* algorithm. In the first experiment, the path planning of a mobile robot when everyone is treated as an ordinary obstacle is studied. In order to achieve this, humans are modeled as circular obstacles, and the path planning algorithm is DWA-A\* algorithm. In the second experiment, the path planning behavior of the robot when modeling the dynamic social area using human status and human group information was investigated. The path planning algorithm used the DWA-A\* algorithm. In order to verify the effectiveness of the proposed algorithm, in the third experiment, the social interaction space is modeled using human status and human group information, and the path planning algorithm uses the DWA algorithm. The experimental results are shown in Fig. 6a. The black line, red line and blue line are the results of experiments 1, 2 and 3, respectively. The PCI and GCI values along the robot trajectory are shown in Fig. 6b, c, respectively.



**Fig. 6** Social interaction space test results

As shown in Fig. 6a, the yellow square and the red five-pointed star represent the starting point and the ending point, respectively. Path1 is the path planned by the social interaction space model and DWA-A\* algorithm, and path2 is the circular obstacle model and DWA-A\*. The path planned by the algorithm, path3 is the path planned by the social interaction space model and the DWA algorithm. As shown in Fig. 6b, c, although the PCI value of path2 is lower than the threshold  $T_c = 0.9$ , the GCI value is higher than the threshold  $T_p = 0.7$ . The PCI value and GCI value of path1 are both lower than the threshold. Compared with path3, path2 has a shorter path. This result shows that the social interaction model can avoid the robot from passing through the interacting crowd, but bypass the crowd, thereby improving the human inner comfort. On the premise of considering human comfort, the path length planned by DWA-A\* algorithm is the shortest, that is, the path planned by DWA-A\* is better than the path planned by DWA.

## 5 Conclusion

This paper studies the navigation of mobile robots in a human–machine environment, and aims to solve the problem that robot behavior should satisfy human inner comfort.

The human interaction space is modeled with a nonlinear scale probability function. The 2D Gauss-Gumbel spatial density function was used to describe the personal space and social interaction space according to the individual state (location, direction, and velocity) of personal information and social group information, respectively. The personal space and social interaction space constitute a Dynamic Social Space (DSS), which considers the psychological comfort of humans. On the basis of the comfort distance of waypoints proposed by Hall, this paper proposes two intuitive metrics to measure human comfort—Personal Comfort Index (PCI) and Group Comfort Index (GCI). The validity of the structure of social interaction space is verified.

**Next Step** The proposed model will be experimented with a real robot. In addition, the next step can be to model people in different postures (sitting or squatting). Social interaction space modeling currently only models the interaction space between people, and the next step can be based on the accurate judgment of human intentions to increase the interaction space modeling between people and things.

## References

1. Chen, W., Xu, J., Zhao, X., et al.: Separated sonar localization system for indoor robot navigation. *IEEE Trans. Ind. Electron.* **99**, 1–1 (2020)
2. Ma, X., Zhou, J., Zhang, X., et al.: Design of a new catheter operating system for the surgical robot. *Appl. Bionics Biomech.* **2021**(1), 1–9 (2021)

3. Yang, H., Qi, J., Miao, Y., et al.: A New Robot Navigation algorithm based on a double-layer ant algorithm and trajectory optimization. *IEEE Trans. Ind. Electron.* **66**(11), 8557–8566 (2019)
4. Song, H., Li, A., Wang, T., et al.: Multimodal deep reinforcement learning with auxiliary task for obstacle avoidance of indoor mobile robot. *Sensors* **21**(4), 1363 (2021)
5. Crew, B.: A closer look at a revered robot. *Nature* **580**(7804), S5–S7 (2020)
6. Malone, N., Chiang, H.T., Lesser, K., et al.: Hybrid dynamic moving obstacle avoidance using a stochastic reachable set-based potential field. *IEEE Trans. Robot.* 1–15 (2015)
7. Song, B., Wang, Z., Zou, L.: An improved PSO algorithm for smooth path planning of mobile robots using continuous high-degree Bezier curve. *Appl. Soft Comput.* **100**(1), 106960 (2021)
8. Rogers, A.J., Hamity, C., Sharp, A.L., et al.: Patients' attitudes and perceptions regarding social needs screening and navigation: multi-site survey in a large integrated health system. *J. General Internal Med.* **35**(16) (2020)
9. Papadakis, P., Spalanzani, A., Laugier, C.: Social mapping of human-populated environments by implicit function learning. In: The IEEE/RSJ International Conference on Intelligent Robots & Systems (2014)

# Fool a Hashing-Based Video Retrieval System by Perturbing the Last 8 Frames of a Video



Chao Hu , Liang Huang , and Ronghua Shi

**Abstract** Studies on adversarial attack have brought people's attention to the safety of deep neural networks (DNNs). Sparse adversarial attack, which is more dangerous than dense adversarial attack, can fool a threat model with a low amount of pixels and perceptibility. However, sparse adversarial attack has not been done extensively on video hashing retrieval. We propose a method to craft sparse adversarial videos on deep hashing retrieval by adding temporal masks on video frames. Adversarial perturbation produces propagation during the video adversarial attack. To study the propagation of sparse adversarial perturbation in video hashing in depth, we develop a cosine similarity curve to show the difference between adversarial video frames and clean video frames. The results show that the perturbation can only propagate from front to back. In addition, to exclude the propagation of perturbations, we conduct experiments to only perturb the last few frames in order to analyze the influence of sparsity on the results. The experimental results show that even when there is no propagation, perturbing the last eight frames can significantly show the ability of adversarial attack to video hash retrieval model. We propose the first targeted white-box sparse adversarial attack on hashing-based video retrieval.

**Keywords** Hashing · Video retrieval · Sparse attack · Propagation

## 1 Introduction

Nowadays, hashing networks are widely applied in content-based retrieval to efficiently convert high-dimensional data into semantic low-dimensional hash code. Video hashing is a hot topic [1, 2], but research on video hashing is difficult due to

---

C. Hu · L. Huang · R. Shi ()

School of Computer Science and Engineering, Central South University, Changsha 410083, China  
e-mail: [shirh@csu.edu.cn](mailto:shirh@csu.edu.cn)

C. Hu

Big Date Institute, Central South University, Changsha 410083, China

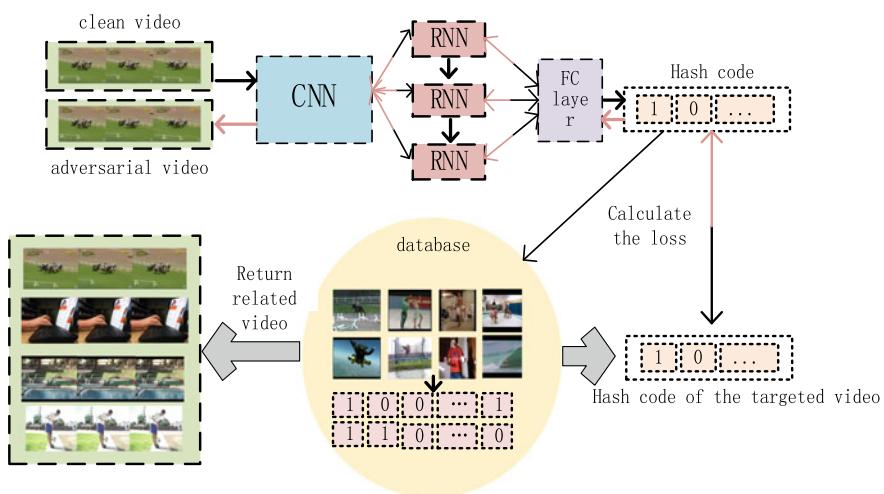
Key Laboratory of Medical Information Research (Central South University), Changsha 410083,  
China

the complicated temporal and spatial characteristics of videos. Research on video hashing has made substantial breakthroughs with deep neural networks (DNNs). Learning-based video hashing methods such as Self-supervised video hashing and Self-supervised temporal hashing [3, 4] can simultaneously preserve the temporal and spatial characteristics of video.

In recent years, DNNs have been proven to be vulnerable to minor perturbations [5]. A lot of studies have been done on adversarial attacks on images [6–8], audio [9], and videos [10–12]. Some studies have shown that adversarial attacks have positive effects. The use of adversarial examples as training examples can contribute greatly to improving the robustness of the network [13]. Research on adversarial attack can also promote human understanding of DNNs. Therefore, it is meaningful to study adversarial attacks.

Adversarial attack can be divided into dense and sparse adversarial attack. Dense adversarial attack has been used to make a great deal of progress on video tasks. However, little work has been done on sparse video adversarial attack.

For these reasons, we decided to study sparse adversarial attack based on video hashing retrieval. After calculating the gradient, we set the projected gradient descent (PGD) as our optimizer.  $\ell_{2,1}$ -norm and  $\ell_2$ -norm are used to constraint the loss function. We also set temporal masks on video frames to control the sparsity of the adversarial attack. Our method can generate adversarial videos with imperceptible perturbations to fool the targeted model. Figure 1 shows the query process of videos and the procedure for creating adversarial examples. We also draw the cosine similarity curve for the adversarial and targeted video frames to study the direction of propagation in depth. We find that the adversarial perturbation can only propagate from front to back. When we perturbate the last frames, the performance of the attack is outstanding even when there is no propagation.



**Fig. 1** The process of querying videos and the creating targeted adversarial videos

Our contributions are as follows:

- As far as we know, this is the first study to create sparse adversarial examples on video hashing retrieval. We have experimentally studied the influence of quantity and position of perturbed frames.
- We have analyzed the propagation direction of adversarial perturbations in video hashing retrieval and have studied the influence of sparsity by experimentally perturbing the last few frames in a video.

## 2 Related Work

Many optimization algorithms have been applied on adversarial attack. Szegedy et al. [14] proposed L-BFGS to generate an adversarial example. Goodfellow et al. [5] proposed the fast gradient sign method (FGSM) to generate adversarial examples with the gradient of the targeted function. Nicolas et al. [8] proposed the Jacobian-based saliency map attack (JSMA), which is based on the gradient of the output layer. Kurakin et al. [7] proposed the projected gradient descent (PGD) to generate adversarial examples with stable adversarial performance.

A lot of research has been done on adversarial attacks of images and videos. Zhao et al. [15] exploited generative adversarial networks (GANS) to craft adversarial images by perturbing visually insignificant areas. Yang et al. [16] introduced tanh as the activation function to solve the problem of gradient disappearance. Wei et al. [12] misclassified videos by exploiting the sparsity and propagation of video perturbation with temporal masks. Chen et al. [10] added adversarial frames to the targeted video to reach the attack goal. Researchers have also used the generative adversarial networks to craft real-time adversarial videos [11] and studied hashing-based video retrieval using a voting mechanism [17].

However, little work has been done on adversarial video hashing retrieval. In this study, we generate sparse adversarial videos on hashing retrieval and study the propagation of adversarial perturbations, which have never been explored.

## 3 Methods

In this section, the problem definition, objective function, and optimization process will be discussed in detail. Temporal masks will be introduced to enable the manual control of sparse attacks.

### 3.1 Problem Definition

For a given query video  $x$ , the hash code  $b$  is generated in a video hash model  $F(,)$ , and the formula for generating the hash code is as follows:

$$b = F(x), b \in \{0, -1\}^L \quad (1)$$

$L$  represents the length of hash code,  $y$  is a video with the targeted label, and  $F(y)$  is the hash code of the targeted video  $y$ . The principle of generating an adversarial video  $\hat{x}$  is minimizing the hamming distance between  $F(y)$  and  $F(\hat{x})$ .

Generally, there are two goals when generating adversarial videos. The first is to make the difference which is hard for naked eyes to distinguish between the query video  $x$  and the adversarial video  $\hat{x}$ . Another goal is to make the result of generating the adversarial video  $\hat{x}$  in the classification model the same as that of the targeted video  $y$ . Unlike video classification that only returns one label, video hash retrieval returns a couple of videos with mean average precision (MAP). Therefore, we define a new goal, which is to reduce the MAP associated with the ground true label and increase the targeted mean average precision (t-MAP). We also add temporal masks on some frames to ensure that certain frames are not perturbed.

### 3.2 Formulation of Overall Objectives

Given  $x$  as the query video,  $\hat{x}$  as the adversarial video, and  $y$  as the target video,  $F(,)$  is the video hash model used to generate the corresponding hash code of the video. The targeted video  $y$  is stored in the database in the form of a hash code  $F(y)$ .  $l(,)$  is the loss function used to calculate the difference between the hash code adversarial video  $\hat{x}$  and the targeted video  $y$ . We use  $l_2$ -norm and  $l_{2,1}$ -norm to constrain the added perturbation  $\varepsilon$ ,  $\varepsilon = \hat{x} - x$ ;  $\lambda$  is a metric to balance  $\varepsilon$  and  $x$ . Hence, the formula of the overall objectives is as follows:

$$\begin{aligned} & \text{argmin } l(F(x + \varepsilon) - F(y)) + \lambda \cdot \|\varepsilon\| \\ & \text{s.t. } \|\varepsilon + x\| \in [0, 1] \end{aligned} \quad (2)$$

In hashing-based retrieval systems, we set hamming distance to measure the distance between hash codes. However, the hamming curve is non-convex and non-smooth [11], which is not suitable for the calculation of the reverse gradient. Therefore, we use the inner product instead of the hamming distance to make the loss function easy to differentiate. We use the tanh function, which can maintain the range of the inner product value at  $(-1, 1)$ , as the activation function of the hashing-based retrieval model. We also introduce  $\alpha$  to solve the problem of gradient disappearance. The final objective function is shown below.

$$\begin{aligned} & \operatorname{argmin} - \langle \tanh \alpha F(x + \varepsilon), F(y) \rangle + \lambda \cdot \|\varepsilon\| \\ & \text{s.t. } \|\varepsilon + x\| \in [0, 1] \end{aligned} \quad (3)$$

For sparse attacks, the temporal mask,  $m$ , is introduced.  $m$  is a matrix,  $m \in \{0, 1\}^b$ , and  $b$  is the number of video frames. Each  $m_t$  corresponds to a video frame,  $x_t$ , where  $x_t$  is the  $t$ -th frame in the video. When  $m_t$  is equal to 0, the corresponding video frame  $x_t$  can no longer be perturbed. The formula of the sparse attack can be expressed in the following form:

$$\begin{aligned} & \operatorname{argmin} \langle \tanh \alpha F(x + m\varepsilon), F(y) \rangle + \lambda \cdot \|m\varepsilon\| \\ & \text{s.t. } \|\varepsilon + mx\| \in [0, 1] \end{aligned} \quad (4)$$

### 3.3 Optimization Process

We will describe the training process in detail. The core idea is to compute the minimum perturbation by gradient descent. We will achieve our goal through multiple iterations. In our experiment,  $\alpha$  will be tuned in a training process. Equation (3) is easy to solve in a neural network using a gradient optimized algorithm (e.g., PGD) because only one variable is present in the equation. We will obtain the convergency value after some iterations. The complete entailed in Algorithm 1 is shown below.

---

**Algorithm 1** Sparse adversarial attack

---

**Require:**

 Query video  $x$ ;

 Targeted video  $y$ ;

 Hashing-based video retrieval model  $F(\cdot)$ ;

Gradient optimized algorithm PGD;

Retrieval algorithm Faiss;

**Adversarial video  $\hat{x}$** 
**Ensure:**
**1: Initialize the input video  $x$** 
**2: Add a perturbation  $m * \varepsilon$  to video  $x$** 
**3: Calculate the hash code  $F(x + m * \varepsilon)$** 
**4: Calculate the loss function  $L(x)$  between  $F(x + m * \varepsilon)$  and  $F(y)$** 
**5: Calculate the gradient  $G(x)$  for  $x$  through loss value  $L(x)$** 
**6: Calculate adversarial perturbation  $\varepsilon$  and update  $\hat{x}$** 
**7: return  $\hat{x}$** 


---

## 4 Experiment

### 4.1 Dataset and Metrics

**Dataset** We use UCF101 [11] as the dataset for our experiment. UCF101 is one of the largest collections of human actions downloaded from YouTube and includes 13,320 video clips worth 27 h. It contains 101 classes of human actions such as haircutting, sky diving, and surfing. We extract 10,000 video clips as the training data and use 3300 video clips for the database in the retrieval system. The remaining 20 video clips are used as query videos.

**Metrics** (1) t-MAP. The t-MAP [10] label is set as the targeted video label in the retrieval process. The t-MAP is used to measure the average accuracy of targeted attacks in a retrieval system. (2) Propagation. Propagation is the ability of adversarial perturbation to propagate from perturbed frames to other frames during the generation of the adversarial examples. (3) Transferability. Transferability is a metric used to measure the precision of adversarial examples under different experimental conditions.

### 4.2 Implementation

In this paper, we use both ResNet152 and long-short term memory (LSTM) as a hash model to generate the hash code of videos. ResNet152, which is a feature extractor, can convert videos into their one-dimensional features. LSTM takes the feature output from ResNet152 as an input vector. The last layer of the fully connected layer is adjusted according to the length of the hash code. We use the PGD method as the optimization algorithm for updating perturbation and use the Faiss method for retrieval. All work is done using Pytorch 1.2.0.

### 4.3 Experimental Analysis

**Tense Adversarial Attack** To verify the effect of generating the adversarial examples on a video hashing system, we conduct four experiments using different hash code lengths. We also conduct experiments based on  $l_2$  and  $l_{2,1}$ -norm constraints. Table 1 shows the results of the experiments.

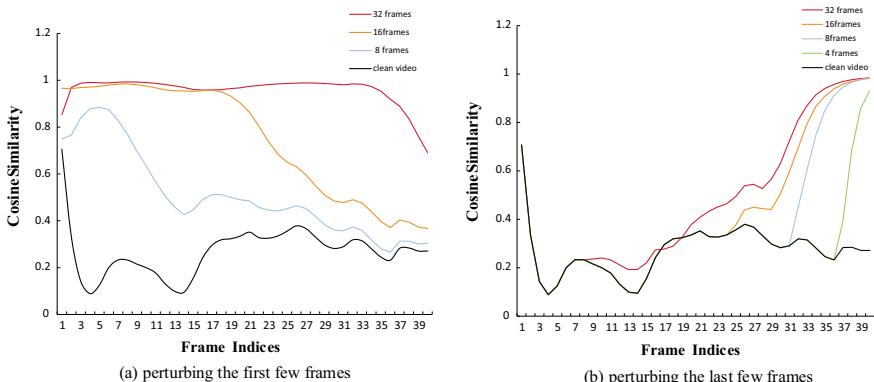
The MAP and t-MAP of the 64-bit hash code are the highest compared to those of other hash code lengths. The t-MAP under 64-bits is very close to the MAP of clean videos, which shows that our method performs well on adversarial attack.

**Table 1** The MAP of a clean video query and t-MAP of an adversarial video query based on  $l_{2,1}$ -norm and  $l_2$ -norm with different hash code lengths

Hash code lengths (bits)	Query video MAP	Adversarial video t-MAP	
		$l_{2,1}$ norm	$l_2$ norm
16	0.418	0.347	0.380
32	0.512	0.401	0.512
64	0.539	0.524	0.530
128	0.409	0.356	0.372

**Propagation** To study the propagation direction of adversarial perturbation, we conduct two separate experiments to perturb the first few frames and the last few frames. We set  $l_{2,1}$ -norm as the constraint function. Classification returns only one label, but multiple related videos and labels are returned after retrieval, and this is not conducive for directly determining the propagation direction of adversarial perturbation. Therefore, we draw cosine similarity curves between adversarial video frames and targeted video frames. The cosine similarity curves are shown in Fig. 2. The hash code length is set to 64-bits.

Figure 2a shows that the cosine similarity curves of frames that are not perturbed are not coincident with those of clean video frames. This means that perturbation occurs on the rear frames. We conclude that adversarial perturbation can propagate from front to back, but the perturbation cannot propagate forever because the differences in cosine similarity between the frames of the adversarial video and targeted video gradually decrease over time. However, when the last few frames are perturbed, the cosine similarity curves of frames that are not perturbed are coincident with those of clean video frames (Fig. 2b). This means that perturbation only occurs in frames



**Fig. 2** Cosine similarity curves of frames in the adversarial video and targeted video under the constraint of  $l_{2,1}$  norm

**Table 2** The results of perturbing the last few frames

$M$	4	8	16	32
t-map	0.469	0.561	0.553	0.516

**Table 3** Results of perturbing the last eight frames of different video categories based on  $l_{2,1}$  and  $l_2$  norms

Different kinds of video	Long Jump		Writing on board		Military parade	
	$l_{2,1}$	$l_2$	$l_{2,1}$	$l_2$	$l_{2,1}$	$l_2$
t-map	0.523	0.535	0.521	0.549	0.519	0.522

that are set to be perturbed. Therefore, we can infer from Fig. 2 that adversarial perturbation can only propagate from front to back.

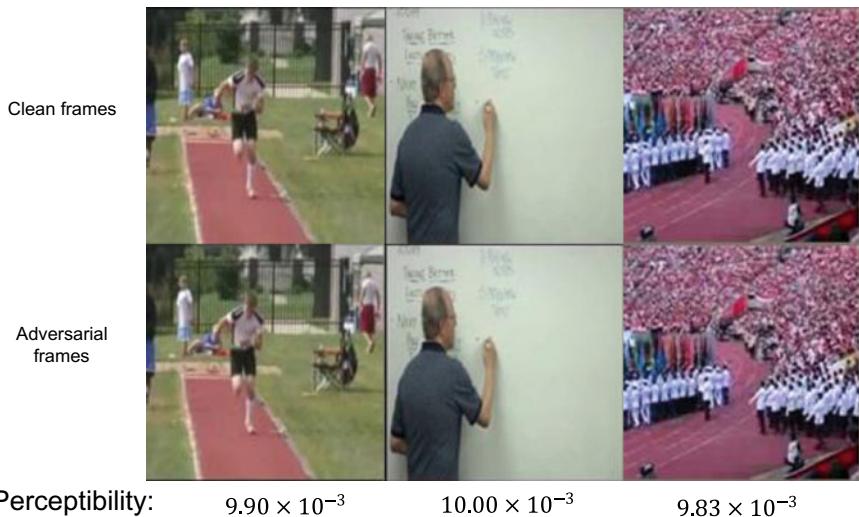
**Sparse Adversarial Attack** Sparse adversarial attack can be divided into temporal sparse and spatial sparse adversarial attack. In this paper, we study only temporal sparse attack. In order to study the effect of the quantity of perturbed frames on the experimental results of sparse adversarial attack, we set up the following experiments. In order to eliminate the influence of propagation, we only perturb the last few frames. The results are shown in Table 2.  $M$  represents the number of perturbed frames.

When the last four frames are perturbed, the t-map is the lowest out of all the other experiments. This may be because the number of perturbed frames is not enough, and hence, the t-map is low. When the last eight frames are perturbed, the t-map is the highest, and this figure is similar to that for tense adversarial attack. This shows that t-map performs well when the last few frames are perturbed. In order to rule out contingency, we conduct three experiments on different video categories under the constraint of the  $l_2$ -norm and  $l_{2,1}$ -norm. The results are shown in Table 3. All the t-maps perform well when the last eight frames are perturbed.

Figure 3 shows visual examples of adversarial videos. Although the maximum perceptibility is  $10.00 \times 10^{-3}$ , we still cannot visually detect the difference between the clean frames and the perturbed frames. Our experiment has a great visual performance.

## 5 Conclusion

We generate sparse adversarial videos under the hashing-based retrieval system and use temporal masks to control the intensity and location of sparsity. From the experiments, we conclude that adversarial perturbation can only propagate backward. When the last eight frames of a video are perturbed, t-map has great performance. We cannot visually detect the difference between the clean frames and the perturbed frames. The position and number of adversarial frames have different effects on experiments.



**Fig. 3** Three visualization examples of the sparse adversarial attack. The frames on the top row are clean frames, and frames on the bottom row are perturbed

However, our experiment only sets the perturbed frame manually. Future studies should be done to calculate the impact of each perturbed frame on the experiment, and we plan to explore the reasons why the adversarial perturbed frames at different positions have different effects on the experimental results.

**Acknowledgements** This work is supported in part by the Humanities and Social Sciences Research Youth Fund Project of the Ministry of Education under Grant 17YJC880037, the National Natural Science Foundation of China under Grant 61977062, and the High Performance Computing Center of Central South University.

## References

1. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4242–4251 (2018)
2. Zhang, X., Lai, H., Feng, J.: Attention-aware deep adversarial hashing for cross-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 591–606 (2018)
3. Song, J., Zhang, H., Li, X., Gao, L., Wang, M., Hong, R.: Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Trans. Image Process.* **27**(7), 3210–3221 (2018)
4. Zhang, H., Wang, M., Hong, R., Chua, T.S.: Play and rewind: optimizing binary representations of videos by self-supervised temporal hashing. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 781–790 (2016)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. [arXiv:1412.6572](https://arxiv.org/abs/1412.6572) (2014)

6. Feng, Y., Chen, B., Dai, T., Xia, S.: Adversarial attack on deep product quantization network for image retrieval. [arXiv:2002.11374](https://arxiv.org/abs/2002.11374) (2020)
7. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. [arXiv: 1607.02533](https://arxiv.org/abs/1607.02533) (2016)
8. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
9. Li, J., Qu, S., Li, X., Szurley, J., Kolter, J. Z., Metze, F.: Adversarial music: real world audio adversary against wake-word detection system. In: Advances in Neural Information Processing Systems, pp. 11931–11941 (2019)
10. Chen, Z., Xie, L., Pang, S., He, Y., Tian, Q.: Appending adversarial frames for universal video attack. [arXiv:1912.04538](https://arxiv.org/abs/1912.04538) (2018)
11. Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S.V., Chowdhury, R., Swami, A.: Adversarial perturbations against real-time video classification systems. [arXiv:1807.00458](https://arxiv.org/abs/1807.00458) (2018)
12. Wei, X., Zhu, J., Yuan, S., Su, H.: Sparse adversarial perturbations for videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8973–8980 (2019)
13. Lin, K-Y., Wang, G.: Hallucinated-IQA: no-reference image quality assessment via adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 732–741 (2018)
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. [arXiv:1312.6199](https://arxiv.org/abs/1312.6199) (2013)
15. Zhao, G., Zhang, M., et al.: Unsupervised adversarial attacks on deep feature-based retrieval with GAN. [arXiv:1907.05793](https://arxiv.org/abs/1907.05793) (2019)
16. Yang, E., Liu, T., Deng, C., Tao, D.: Adversarial examples for hamming space search. IEEE Trans. Cybern. (2018)
17. Bai, J., Chen, B., Li, Y., Wu, D., Guo, W., Xia, S-t, Yang, E-h.: Targeted attack for deep hashing based retrieval. [arXiv:2004.07955](https://arxiv.org/abs/2004.07955) (2020)

# Multi-level Road Damage Identification Algorithm Based on Vehicle-Mounted Smartphone



Deng Ma<sup>✉</sup>, Kai Gao<sup>✉</sup>, and Ronghua Du<sup>✉</sup>

**Abstract** In order to improve the detection timeliness of abnormal road surfaces and reduce the detection cost, this paper provides a road damage identification algorithm based on the vehicle-mounted smartphone. The algorithm uses a vehicle-mounted smartphone to detect whether the vertical acceleration exceeds the set threshold as the direct condition of the road damage degree. An optimized Butterworth filter was designed to denoise the vertical acceleration of the collected vehicles. Then, an improved Gaussian background model was used to select the appropriate threshold conditions, and the multi-grade classification of abnormal road damage was realized based on the support vector machine algorithm. The experimental results show that the classification accuracy of the algorithm can reach 92.34%. Compared with the existing detection methods, the proposed method has lower detection costs and realizes the fine classification of road damage, which is convenient for the road maintenance department to take reasonable measures in time.

**Keywords** Road damage identification · Vertical acceleration · Smartphones · SVM

## 1 Introduction

With the rapid development of China's social economy, the modernization level of urban construction has been continuously improved, and the road traffic network has been improved year by year. By the end of 2019, China's total highway length had increased to 5.0125 million km, with 672,000 km of roads at grade II or above, including 149,600 km of expressways. In the same year, the highway maintenance mileage was 4.9531 million km, accounting for 98.8% of the total highway mileage, an increase of 0.6% compared with 2018 [1]. As an essential part of the transportation system, the highway has made significant contributions to the stable and healthy

---

D. Ma · K. Gao (✉) · R. Du  
Changsha University of Science & Technology, Changsha 410114, China  
e-mail: [kai\\_g@csust.edu.cn](mailto:kai_g@csust.edu.cn)

development of the economy. However, in the current social and economic sustainable development in the background, the highway transportation under pressure is more and more significant, plus the highway investment must be diversified, maintenance funding imbalance problems, such as highway maintenance demand in China is not very good for a long time to meet with the development of science and technology, happening in our country existing highway maintenance pattern changes with each passing day, The information construction of highway maintenance industry has become an important trend of future development.

Nowadays, the smartphone has become a necessity in People's Daily life. It can obtain the current position, acceleration, and other mobile phone states as an intelligent terminal with numerical calculation and external perception. Because of its rich sensors, more and more scholars in transportation are trying to apply smartphone sensors to road disease detection and other aspects [2]. This paper designed a kind of abnormal road recognition algorithm based on acceleration by calling the smartphone triaxial acceleration sensors in the vibration signal and use the Butterworth filter to filter noise, after SVM algorithm to the pretreatment of data classification, and then implements the abnormal damage of road surface is different degree of recognition.

## 2 Foreign and Domestic Research Background

Domestic and foreign experts and scholars have conducted long-term research on abnormal pavement identification technology. The existing abnormal pavement identification methods mainly include three kinds: based on 3D reconstruction, based on visual recognition, and based on acceleration [3–10]. Koch proposed a detection method based on image recognition, which can automatically detect pits and grooves in asphalt pavement [6]. The method uses image segmentation to compare the texture of the defect area and non-defect area to realize the recognition of abnormal road surface. However, Koch's method could not effectively count the width and depth of abnormal road surfaces. Jog proposed a recognition method combining two-dimensional image recognition and three-dimensional reconstruction based on Koch, which could identify abnormal road surfaces and measure the number, width, and depth of abnormal road surfaces [7]. For improved accuracy of the collected abnormal road surface, Aki used a three-dimensional laser scanner to reflect laser pulses and obtain accurate three-dimensional point clouds to establish a mathematical model [8].

Chen acquires vehicle acceleration, speed, and position information by installing some acceleration sensors and GPS modules on the vehicle and uses a Gaussian background model to identify abnormal road surfaces [9]. In order to adapt the abnormal road recognition method to different vehicle speeds, Harikrishnan proposed an improvement on Gaussian background based on Chen [10]. In order to improve intelligent detection. Zhang Jinxi collected vehicle vibration information by installing a three-axis accelerometer on the vehicle, extracted characteristic values by wavelet

denoising, and then explored the relationship between signal characteristic values and flatness by using the GA-BP neural network [11]. However, this method needs much computation, and the interpretability of the neural network model is inadequate.

With the rapid development of smartphones, accelerometers, gyroscopes, and other sensor modules equipped with them have become essential tools for detecting abnormal road surfaces. Luis C used the accelerometer and gyroscope sensors in smartphones to obtain the required acceleration and other data, compared the recognition accuracy of various machine learning classification algorithms, and finally selected the gradient-aided optimization strategy and random forest algorithm to identify abnormal road surfaces [12]. Yi proposed an anomaly index algorithm based on smartphones to monitor the road surface, but this algorithm has a poor effect on recognizing abnormal road surfaces without speed bumps [13].

Domestic and foreign scholars have made specific achievements in road condition identification methods, among which two methods based on visual identification and three-dimensional reconstruction have better identification effects. However, these two methods have high equipment and high-cost requirements, which are not conducive to popularization and use. Most of the abnormal road surface detection methods based on acceleration are relatively single, and many methods need to install accelerometers, gyroscopes, and other sensors on the vehicle. At present, with the popularity of smartphones, its integrated acceleration sensor, positioning system, mobile network, and substantial computing and processing ability provide necessary conditions for its application in vehicle road detection.

### 3 Abnormal Road Surface Recognition Algorithm

#### 3.1 Filter Design

A random noise will be generated by the difference of road conditions in-vehicle driving, and the vibration of the vehicle body will also produce some noise. In order to eliminate the influence of these factors, Butterworth's seventh-order low-pass filter is selected in this paper for noise elimination.

Butterworth filter is an electronic filter, which does not produce ripples in the passband, and the response of frequency curve does not fluctuate in the passband, and the response of the frequency in the stopband gradually drops to zero. Compared with other kinds of filters, the Butterworth filter has the characteristics of maximum flatness, and the processed result is smoother [14].

Butterworth lowpass filter has five technical parameters: order N, passband cutoff frequency  $W_p$ , stopband cutoff frequency  $W_s$ , passband ripple peak  $R_p$ , stopband ripple peak  $R_s$ . The acquisition frequency of triaxial acceleration set in this paper is 400 Hz, so the signal acquisition frequency is set  $F_s = 400$  Hz here. References show that abnormal road signals are usually concentrated in the frequency range 30 Hz, So set the passband cutoff frequency  $W_p = 30$  Hz, stopband cutoff frequency

$W_s = 2W_p = 60$  Hz, passband ripple peak  $R_p = 0.5$  dB, stopband ripple peak  $R_s = 30$  dB. The Butterworth low-pass filter order is calculated by the formula  $N = \lg \sqrt{E} / \lg \frac{W_s}{W_p}$ , where  $E = 10^{\frac{R_s}{10}} - 1/10^{\frac{R_p}{10}} - 1$ .

The signal time domain and frequency domain of the Butterworth low-pass filter denoising after the above Settings are compared with the original signal time domain and frequency domain. It can be seen from Fig. 1 that after filtering and denoising, the high-frequency noise in the signal spectrum is effectively filtered out, and the acceleration spectrum is concentrated in 0–30 Hz. Therefore, Butterworth low pass filter meets the design requirements.

### 3.2 The Threshold Condition

After the Kolmokorov-Smilov test, it is found that the vertical acceleration of the vehicle on the flat road conforms to the Gaussian distribution, so vertical acceleration value of the vehicle will meet Formula 1 [15].

$$\eta(A|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(A-\mu)^2}{2\sigma^2}} \quad (1)$$

Among them:

- A Vertical acceleration of vehicle.
- $\mu$  The expected value of the vertical acceleration.
- $\sigma$  The standard deviation of the vehicle's vertical acceleration.

Because the vertical acceleration generated by the vehicle passing the abnormal road surface does not obey the Gaussian distribution, the vertical acceleration generated by the abnormal road surface should conform to Formula 2.

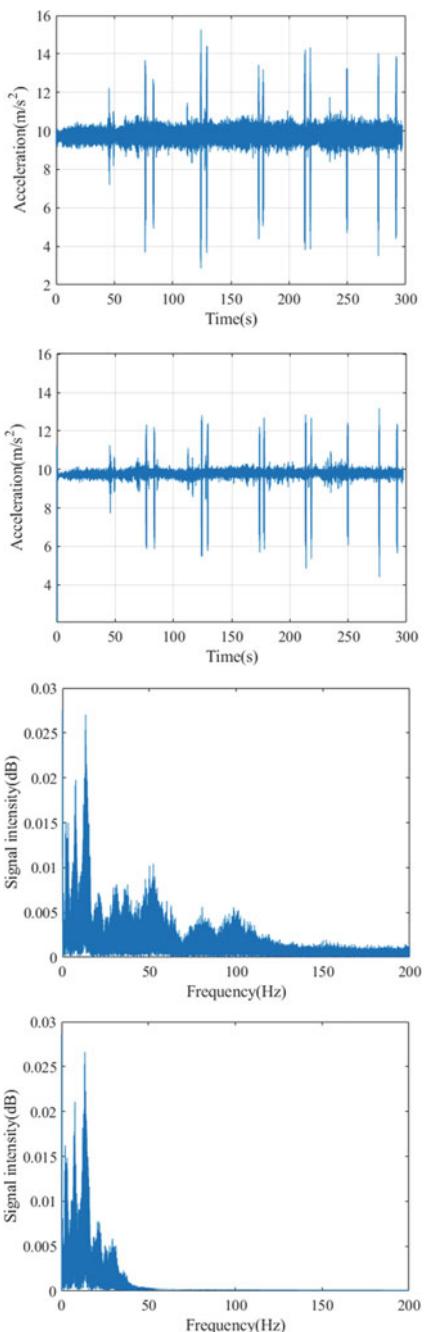
$$\left| \frac{A - \mu}{\sigma} \right| > \frac{v}{T_v} \cdot T_a \cdot \sigma \quad (2)$$

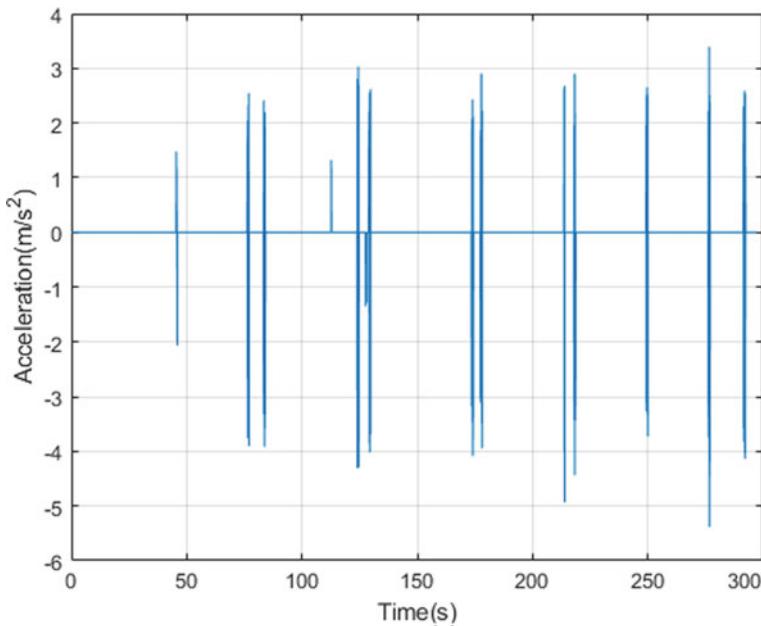
Among them:

- $v$  Vehicle speed.
- $T_v$  Speed threshold.
- $T_a$  Acceleration threshold.

Considering the different suspension system parameters of different models, the speed threshold and acceleration threshold are set as 20 Km/h and 3 m/s<sup>2</sup> respectively in this paper [16, 17]. After the threshold condition is added, the time domain diagram of vertical acceleration is shown in Fig. 2.

**Fig. 1** Time-frequency domain comparison of signals before and after filtering

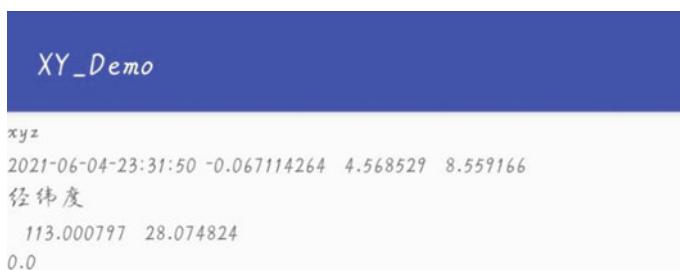




**Fig. 2** Time-domain diagram of vertical acceleration after setting threshold conditions

### 3.3 Data Collection and Partition

At present, most smartphones are equipped with accelerometers and GPS positioning modules [18]. In this paper, a smartphone with the model of Redmi Note8 Pro is used for data collection. In order to collect the original data, we installed an application on the smartphone that can record the time, acceleration, velocity, longitude and latitude coordinates, and other data detected by the smartphone. The page of data collection is shown in Fig. 3. Different types of mobile phone accelerometers may



**Fig. 3** The page of data collection

**Table 1** Size of training set and test set

The road surface type	Training set size	Test set size
Light road surface	1116	272
Moderate road surface	751	178
Serious road surface	219	69
Flat road surface	1150	290
Total	3236	809

have different coordinate systems, but most detection accuracy can meet the required data accuracy.

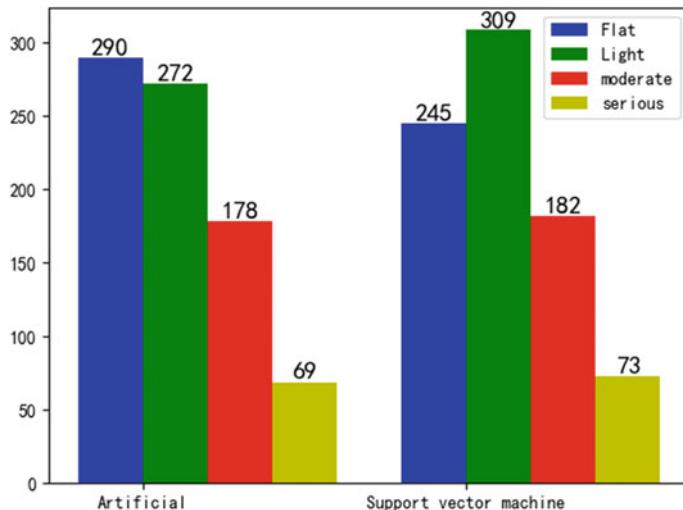
According to the degree of damage of abnormal road surface, the road surface was divided into three types, namely light, moderate, and serious. The light, moderate, and serious road surface were all damaged and the degree of damage increased successively. The degree of damage is divided by a fuzzy logic control algorithm, and threshold conditions are further set. Here, the threshold conditions for light, moderate and severe damage are respectively  $4 \text{ m/s}^2$ ,  $5 \text{ m/s}^2$  and  $6 \text{ m/s}^2$ . Then, 80% of data from the dataset collected by the triaxial acceleration of smart phones were randomly selected and divided into training sets, and the remaining data were divided into test sets. The sizes of training sample sets and test sample sets are shown in Table 1.

### 3.4 Support Vector Machine Multi-classification

SVM is a linear classifier with the most considerable interval defined in the feature space. It is a binary classification model and is often used to solve data classification problems. It belongs to one of the supervised learning algorithms. The basic idea of SVM is to divide the data set correctly and separate the hyperplane with the most considerable geometric interval so that the data points in the training sample set can be as far away from the classification plane as possible [19].

In this paper, Python language is used to complete the parameter setting of the multi-classification support vector machine and input the training set for classification training. Then, the precision test of the trained support vector machine is carried out by using the data in the test set. Among them, the penalty parameter of multi-classification SVM was set as 3, and the kernel function was RBF Gaussian kernel function. The results of the test set classification were shown in Fig. 4, and the accuracy on the test set could reach 92.34%. As can be seen from Fig. 4, SVM can effectively classify road surface types, and its main error comes from the classification of flat road surface and road surface with light damage degree. The classification effect of road surface with moderate damage degree is the best.

Considering the high frequency of data collection, it is difficult to collect data and send it to other terminals for processing. Therefore, to realize the function of real-time detection, this paper adds data processing, abnormal road surface recognition,



**Fig. 4** Comparison of classification results

and other functions on the basis of the application shown in Fig. 3 so that it can complete the identification of abnormal road surface on the mobile terminal.

## 4 Conclusion

In view of the vertical acceleration in the process of vehicle driving, the support vector machine classification is used after pretreatment by filtering noise reduction, adding threshold conditions, and other methods and the following main conclusions are drawn.

1. Using the real-time traffic data collected by the three-axis accelerometer and denoising by Butterworth filter, the vertical acceleration data meeting the requirements can be obtained.
2. The support vector machine algorithm can be used to classify the abnormal pavement types effectively, and its classification accuracy can reach 92.34%.

At present, many scholars in the field of road are devoted to studying the detection of abnormal road surface. Based on the research in this paper, vehicle types and road surface roughness can be further expanded. In addition, intelligent algorithms such as neural networks can be used to improve classification accuracy.

## References

1. Ministry of Transport of the People's Republic of China: In: 2019 Statistical Bulletin on the Development of the Transport Industry. China Communications News (2020) (Chinese)
2. Wahlström, J., Skog, I., Händel, P.: Smartphone-based vehicle telematics: a ten-year anniversary. *IEEE Trans Intell Transport Syst* (2017)
3. Quintana M, Torres J, Menéndez JM (2016) A simplified computer vision system for road surface inspection and maintenance. *17*(3):608–619
4. Ai-min, S.: Recognition and measurement of pavement disasters based on convolutional neural networks. *China J Highway Transport* **31**(01), 1–10 (2018). (Chinese)
5. Li, W., Burrow, M., Li, Z.: Automatic road condition assessment by using point laser sensor. In: 2018 IEEE Sensors, pp. 1–4. IEEE, New Delhi (2018)
6. Koch, C., Brilakis, I.: Pothole detection in asphalt pavement images. *Adv. Eng. Inform.* **25**(3), 507–515 (2011)
7. Jog, G.M., Koch, C., Golpavar-Fard, M., Brilakis, I.: Pothole properties measurement through visual 2D recognition and 3D reconstruction. In: Proceedings of the ASCE International Conference on Computing in Engineering (2012)
8. Aki, M., et al.: Road surface recognition using laser radar for automatic platooning. *IEEE Trans. Intell. Transport. Syst.* (2016)
9. Chen, K., Lu, M., Tan, G., et al.: CRSM: crowdsourcing based road surface monitoring. In: IEEE International Conference on High Performance Computing & Communications & IEEE International Conference on Embedded & Ubiquitous Computing (2014)
10. Varun, P., Gop.: Vehicle vibration signal processing for road surface monitoring. *IEEE Sens J* **17**(16):5192–5197
11. Jinxi, Z.: Research on intelligent detection method of road roughness based on driving vibration. *J China Foreign Highway* **40**(01), 31–36 (2020). (Chinese)
12. Carlos, R., Gonzalez-Gurrola, L., Wahlstrom, J., et al.: Becoming smarter at characterizing potholes and speed bumps from smartphone data—Introducing a second-generation inference problem. *IEEE Trans. Mob. Comput.* **99**, 1–1 (2019)
13. Yi, C.W., Chuang, Y.T., Nian, C.S.: Toward crowdsourcing-based road pavement monitoring by mobile sensing technologies. *IEEE Trans. Intell. Transp. Syst.* **16**(4), 1905–1907 (2015)
14. Dawei, W.: Design of Butterworth analog filter based on Matlab. *Mod. Electron Technol* **35**(21), 71–72, 75 (2012) (Chinese)
15. Jing, F.: Study on crack distribution characteristics of highway tunnel based on K-S test method. *J Zhejiang Inst Commun* **18**(04), 12–16 (2017). (Chinese)
16. Ronghua, D., Gang, Q., Kai, G., Lin, H., Li, L.: Abnormal road surface recognition based on smartphone acceleration sensor. *Sensors* **20**(2), 451 (2020)
17. Varun, P.G.: Vehicle vibration signal processing for road surface monitoring. *IEEE Sens. J.* **17**(16), 5192–5197 (2017)
18. Sensors Overview\_Anyroid Developers: [http://developer.android.com/guide/topics/sensors/sensors\\_overview.html](http://developer.android.com/guide/topics/sensors/sensors_overview.html). Last accessed 2014
19. Xiaoming, X.: SVM Parameter Optimization and Its Application in the Classification. Dalian Maritime University (2014) (Chinese)

# A Digital Twin Model for Battery Management Systems: Concepts, Algorithms, and Platforms



Mi Zhou , Lu Bai , Jiaxuan Lei , Yibin Wang , and Heng Li

**Abstract** In this paper, we propose a digital twin model for battery management systems (BMS). We first discuss the corresponding concepts about the digital twin model of battery management systems. Then, the state-of-charge (SoC) and state-of-health (SoH) estimation algorithms are presented in an integrated fashion for the monitoring and prognostics. Concretely, the extended Kalman filter algorithm (EKF) is used in this paper for the estimation of SoC, which improves the robustness of digital twin model, and the particle swarm optimization algorithm (PSO) is used in this paper for the estimation of SoH. The embedded system platforms are introduced to implement the proposed digital twin model. In the end of this paper, by using the experimental data obtained from the actual circuit experiment and using the Simulink module of MATLAB to simulate the digital twin model proposed in this paper, we verified that the digital twin model proposed in this paper for BMS has good performance in the Gaussian white noise condition.

**Keywords** Battery · Digital twin · SoC · SoH · BMS

## 1 Introduction

Secondary batteries play an extremely important role in the emerging power and energy systems, e.g., smart grid and electric vehicles, where batteries can be discharged to support the load or charged to store the excessive energy [1]. Dominated secondary batteries in the market include Lead-Acid batteries, Li-ion batteries, and supercapacitors, where each of them has different applications, e.g., Lead-Acid batteries have been utilized in the automotive industry for the starting, lighting and ignition (SLI) purposes, and Li-ion batteries are popular in the electric vehicles, and supercapacitors are mostly used in fast/discharging application scenarios [2].

---

This work is supported by National Natural Science Foundation of China (No. 61803394). Mi Zhou, Lu Bai, and Jiaxuan Lei contribute equally to this work.

---

M. Zhou · L. Bai · J. Lei · Y. Wang · H. Li

School of Computer Science and Engineering, Central South University, Changsha, China  
e-mail: [liheng@csu.edu.cn](mailto:liheng@csu.edu.cn)

Battery management systems (BMS) are crucial for the safe and efficient operation of batteries [3]. The BMS is actually an embedded system, where various sensors can be applied to collect the voltage, current and temperature of batteries. The measurements are transmitted to the micro-controller, and based on which the control signal is generated to manage battery cells. The functions of BMS include state measurement and estimation, cell balancing, charging/discharging control [4].

The design of BMS includes two aspects: one is the hardware design and the other is software design [5]. In the hardware design, different physical components need to be analyzed and connected in a logical way, e.g., each cell needs to be connected with corresponding sensors, where the sensor output is connected with the analog-to-digital port of the micro-controller. Similarly, the control signal from the micro-controller needs to drive the actuators (e.g., switches) of the BMS through the electric wire. In the software design, two flows need to be considered: one is the state estimation and the other is the state management. In the state estimation flow, both the state-of-charge (SoC) and state-of-health (SoH) need to be estimated, which is preferred in a collective way [6–9]. In the state management flow, different control algorithms, e.g., cell balancing/charging control/discharging control algorithms are designed to achieve the corresponding purposes [10–12].

Although extensive studies have been conducted on battery management systems for the battery modeling [13], SoC estimation [6, 7], SoH estimation [8, 9], cell balancing [10, 11], charging/discharging control [12], to name a few, existing explorations are still restricted in an ad-hoc way. Actually, in a battery management system, different factors are coupled together, which needs to be considered in a collective way. For instance, the accuracy SoC estimation affects the battery balancing, and the cell balancing control also affects SoC estimation accuracy. Thus, the battery modeling, battery state estimation, and state management of BMS need to be analyzed and designed in a systematic way, which in turn, requires a systematic model to represent the BMS.

A digital twin is digital counterpart of a physical system [14], where the digital counterpart can be used to estimate and predicate the states of the physical system, which can be further used to manage the physical system [15]. In the digital twin framework, the physical part consists of the battery cells, balancing circuits, and additional electrical components; and the digital twin consists of battery modeling, battery state estimation, and battery state management [16–18]. In this sense, if we can build the digital twin model of the battery management system, we can explore the battery modeling, battery state estimation and battery state management in a single model, which provides the insight to designing advanced BMSs.

In this paper, a digital twin model is proposed for battery management systems. We first introduce the concepts of the BMS digital twin. Then, the battery modeling, SoC estimation, SoH estimation algorithms are presented and analyzed in detail. Thereafter, we introduce the digital twin platform used in the performance evaluation. Experiment results verify that the proposed digital twin model can characterize the BMS accurately.

The remainder of this paper is organized as follows. In Sect. 2, we introduce the basic concept of the BMS digital twin. Section 3 presents the digital twin algorithms.

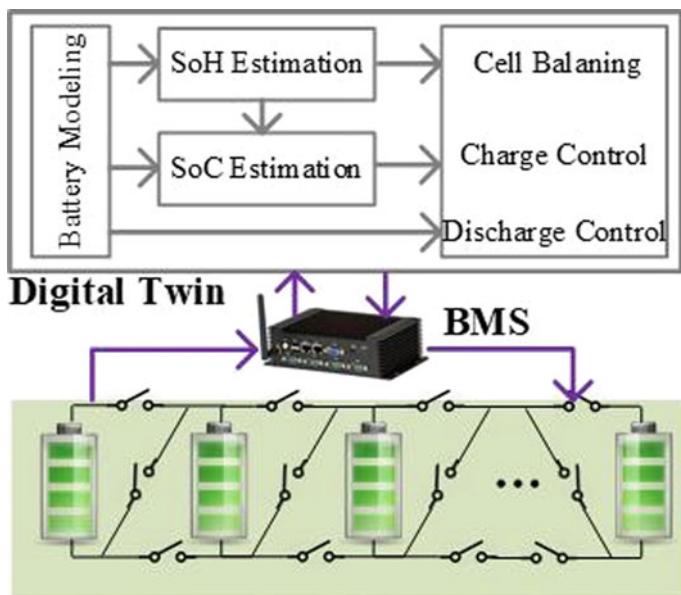
The digital twin model is proposed in Sect. 4. Experiment results are provided in Sect. 5. We conclude the paper in Sect. 6.

## 2 BMS Digital Twin Model

In this section, we introduce some basic concepts about the digital twin model of the battery management systems. As shown in Fig. 1, the whole system consists of three subsystems: battery cells, micro-controller, and the digital twin. The details are introduced as follows.

### 2.1 Battery Cells

In a practical battery storage system, multiple battery cells are connected in series and parallel to satisfy the voltage and power requirement of the application scenario. Additional circuits are typically embedded in battery systems to achieve the balancing of cells. Two categories of balancing circuits can be employed, i.e., passive balancing circuit and active balancing circuit. In the passive balancing circuit, passive components, e.g., resistors or diodes, are connected with cells to dissipate the excessive



energy of high-voltage ones [19]. The energy efficiency of the passive balancing circuit is relatively low, but the circuit benefits from the low cost and small size, which is favored in low power applications. In the active balancing circuit, energy storage units, e.g., inductors or DC-DC converters, are applied to transfer the energy from the high-voltage cells to low-voltage ones [20]. The active balancing circuit benefits from high efficiency, but both the size and cost are high, and thus is typically applied in high power applications. Recently, the reconfigurable battery system has emerged as a new BMS, where the configuration of the battery cells can be adjusted dynamically according to the load requirement [21]. These balancing circuits provide various choices for the designer when designing battery management systems.

## 2.2 Micro-controller

Micro-controller plays a vital role in collecting battery states and in sending control signals to the battery cells. To do that, a correct and logical physical connection is necessary. In the state monitoring flow, corresponding sensors, e.g., voltage/current/temperature sensors are connected to each cell to measure the battery state. The output of the sensors is connected to the analog-to-digital port of micro-controllers, and then are stored therein. With the measured states, the digital twin algorithms can be implemented to model, estimate and control the battery states. In the state management flow, the GPIO port of the micro-controller is connected to the actuator of the BMS, e.g., switch, relay, or DC-DC converter. Different control strategies including switching control, PWM control can be applied to regulate the states of circuits and batteries.

## 2.3 Digital Twin

A digital twin represents the mathematical abstraction of the BMS. Micro-controllers provide a physical platform for the digital twin to be implemented. Recently, with the information technology development, the digital twin model can also be implemented in the cloud server, where the role of micro-controllers becomes a “flow channel” to transmit the battery state to cloud, and send cloud computing result to BMS.

A digital twin model including the algorithms: battery modeling algorithm, SoC estimation algorithm, and SoH estimation algorithm, battery balancing algorithm, and battery charging/discharging control algorithm. In the battery modeling algorithm, the equivalent electrical model of the battery can be applied, which is further discretized and programmed in micro-controller or cloud servers. In the SoC estimation algorithm, classical observer-based approach and the emerging machine learning method can be applied to estimate the SoC of batteries in real time. In the SoH estimation algorithm, different recursive methods can be adopted to estimate the SoH in the long term. In the battery balancing algorithm, feedback control law is

designed to balance the voltage/SoC of batteries based on the designed active or passive balancing circuit. In the charging/discharging control algorithm, a bidirectional DC-DC converter is typically adopted to achieve the energy flow between batteries and the load.

In the following, we emphasize three digital twin algorithms, battery modeling, battery SoC estimation, and battery SoH estimation.

### 3 BMS Digital Twin Algorithms

#### 3.1 Battery Modeling

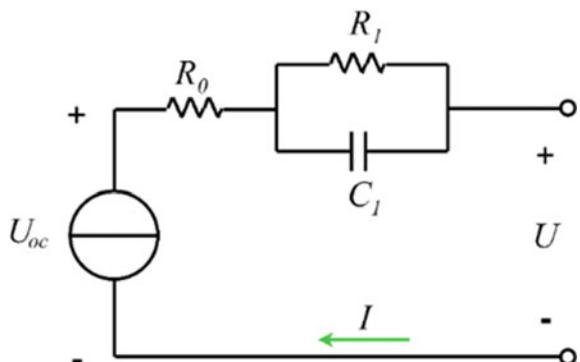
The existing battery models can be usually divided into two categories, i.e., electrochemical model and equivalent circuit model. Among them, the electrochemical model has high precision but many parameters and complex structure, which is not suitable for SoC online estimation scenarios. Neural network model needs a large number of experimental data for learning and training, and needs strong computing ability. In comparison, the equivalent circuit model has fewer parameters and is easy to identify, and has high estimation accuracy. As shown in Fig. 2, the Thevenin model is applied to characterize the battery dynamics:

$$\dot{U}_1 = -\frac{U_1}{R_1 C_1} + \frac{I}{C_1} \quad (1)$$

$$U = U_{oc} - U_1 - IR_0 \quad (2)$$

Thevenin model consists of a voltage source  $U_{oc}$ , resistance  $R_0$  and the parallel network  $R_1 C_1$ .  $U_1$  and  $U$  are the terminal voltage of RC circuit and the terminal voltage of Thevenin model respectively. The structure of this model is relatively

**Fig. 2** Thevenin battery model



simple, the parameters are less and easy to identify, and it can characterize the dynamics of the battery, so it has good practical engineering application value.

### 3.2 SoC Estimation

The EKF was developed on the basis of the Kalman filter, which extends the Kalman filter algorithm to nonlinear Gaussian systems.

Generally, the two main components of Kalman filtering are the prediction part and the update part. As shown in Eqs. (3) and (4) are the prediction equations of the Kalman filter, where  $A$  is the state transition matrix,  $B$  is the input control matrix,  $\hat{x}_k^-$  is a prior estimate of the state,  $\hat{x}_k$  is the posterior estimate of the state,  $u_k$  is the input,  $P_k^-$  is covariance matrix of prior estimation error  $e_k^- = x_k - \hat{x}_k^-$ .  $Q$  is the covariance matrix of process noise  $w_k$ .

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \quad (3)$$

$$P_k^- = AP_{k-1}A^T + Q \quad (4)$$

Correspondingly, the updated equations for Kalman filtering are (5)–(7), Where  $P_k$  is covariance matrix of posterior estimation error  $e_k = x_k - \hat{x}_k$ ,  $Q$  is the covariance matrix of measurement noise  $v_k$ .

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \quad (5)$$

$$\hat{x}_k = \hat{x}_k^- + K_k (y_k - H\hat{x}_k^-) \quad (6)$$

$$P_k = (I - K_k H)P_k^- \quad (7)$$

EKF is applicable to nonlinear systems. The space-state equations of EKF are (8) and (9).

$$x_k = f(x_{k-1}, u_{k-1}) + w_{k-1} \quad (8)$$

$$y_k = g(x_k) + v_k \quad (9)$$

The primary expression of EKF can be derived by linearization of multivariate function with the first-order expansion of Taylor series.

In summary, the  $I$  is selected as the input,  $SoC$  and  $U_1$  as the state, and the terminal voltage  $U$  as the measurement, so the state equation and measurement equation can be expressed as Eqs. (10) and (11):

$$\left\{ \begin{array}{l} SoC(k) = w_{oc}SoC_{oc}(k) + w_{ah}SoC_{ah}(k) \\ U_1(k) = \left( 1 + \frac{1}{R_1(SoC(k-1))C_1(SoC(k-1))} \right)U_1(k-1) \\ \quad + \frac{1}{C_1(k-1)}I(k-1) \end{array} \right. \quad (10)$$

$$U(k) = U_{oc}(SoC(k)) - U_1(k) - I(k)R_0(SoC(k)) \quad (11)$$

where  $T$  is the sampling period, and there are:

$$SoC_{oc}(k) = \frac{U(k) + U_1(k) + I(k)R_0(SoC(k-1)) - \beta_n(k)}{\alpha_n(k)} \quad (12)$$

$$SoC_{ah}(k) = SoC(k-1) - \frac{I(k)T}{Q_{rated}} \quad (13)$$

### 3.3 SoH Estimation

Particle swarm optimization (PSO) can find more suitable parameters by simulating the behavior of the group. It is widely used because of its high adaptability and anti-interference ability. The main flow of particle swarm optimization is as follows: Firstly, the position  $x_i$  and velocity  $v_i$  of all particles are randomly generated and initialized, and the appropriate initial value is determined according to the number of parameters. Then, in the algorithm iteration, the position and velocity are updated according to the best solution of each particle in previous generations and the best solution of all particles in previous generations.

The residual capacity of the battery will decrease during the operation of batteries, resulting in the capacity attenuation of the battery. The SoH of the battery indicates the aging level of the battery. The SoH of a battery can be defined as the ratio of the nominal capacity to the remaining capacity.

Based on the current and voltage measurement data of particle swarm optimization algorithm, according to the fitness function  $f$ , resistance  $R_{0,1}$ , capacitance  $C_1$  and other parameters can be identified.

$$F = \frac{1}{N} \sum_{i=1}^N (U_{t,i} - \hat{U}_{t,i})^2 \quad (14)$$

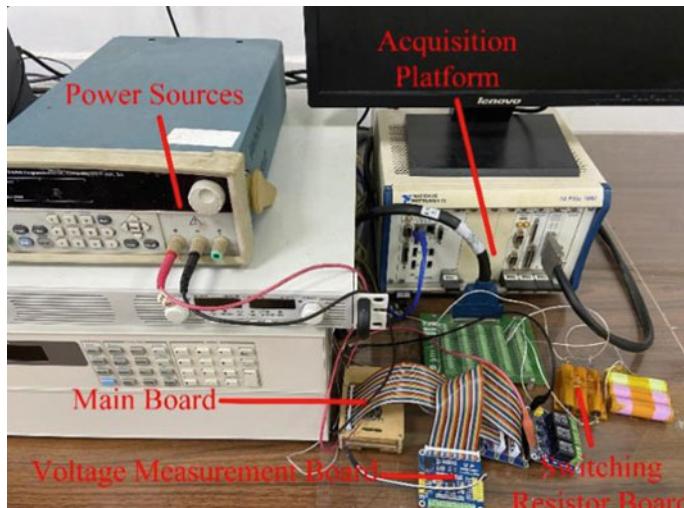
where  $N$  is the number of particles,  $U_{t,i}$  is the voltage of the battery, and  $\hat{U}_{t,i}$  is the estimated battery voltage. When the estimated parameters converge, the algorithm will stop. Then the SoH can be calculated with the identified parameters.

## 4 BMS Digital Twin Platform

As shown in Fig. 3, the hardware platform is composed of four parts: main board, switching resistance board, voltage measurement board, power supply and data acquisition module.

1. Main Board: Considering the weight and price, Raspberry Pi was selected as the main control module. It has a ARM Cortex-A72 CPU with 8 GB LPDDR4 SDRAM, 5.0 Bluetooth, 2.4 GHz and 5 GHz dual-band WiFi and other resources.
2. Switching Resistor Board: Three lithium-ion batteries are parallel connected with the resistor through the corresponding relay channel.
3. Voltage Measurement Board: The ADS1256 chip is utilized with a 8-channel analog-to-digital converter (ADC) and sampling rate of 30 kHz.
4. Power Sources: The power sources supply is divided into two parts. The constant-current power supply charge batteries with a constant current. The DC 24 V power supply provides the operating voltage for the micro-controller and sensors.
5. Acquisition Platform: The data acquisition module includes PXI equipment, measurement board and LabView of upper computer. The PXI platform measures the battery voltages through the measurement board and displays the voltage profiles in the upper computer.

The parameter settings in the circuit are as follows: the rated capacity of the battery is 2600 mAh, the reference voltage  $v_0 = 3.6$  V, the charging current  $i_c = 3.6$  A, balancing resistance  $R = 1 \Omega$ , and sampling period  $T = 0.001$  s.



**Fig. 3** Hardware setting of the digital twin platform

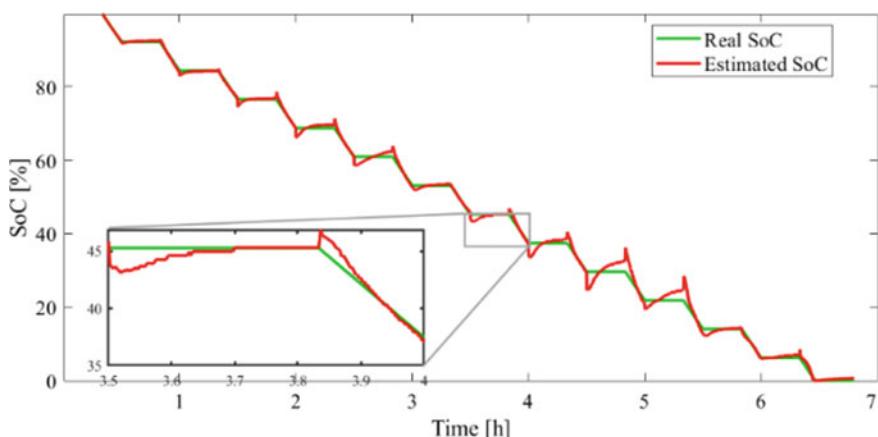
We briefly discuss how the testbed operates during the charging process. As the main control module, Raspberry Pi measures the terminal voltage of each cell by controlling the high-precision voltage sampling module. Through measurement, the controller adjusts the switching state according to the programming algorithm. The measurement data of the battery is sent to the cloud server through HTTPS protocol and stored in the database for data visualization in the Web application. HTTPS protocol adds SSL layer on top of TCP/IP model. The client encrypts the data and sends it to the server. The server obtains the data after decryption, which ensures the security and privacy. On the Raspberry Pi operating system, Python development and software programs are used to convert the cloud control signals into physical values.

## 5 BMS Digital Twin Evaluation

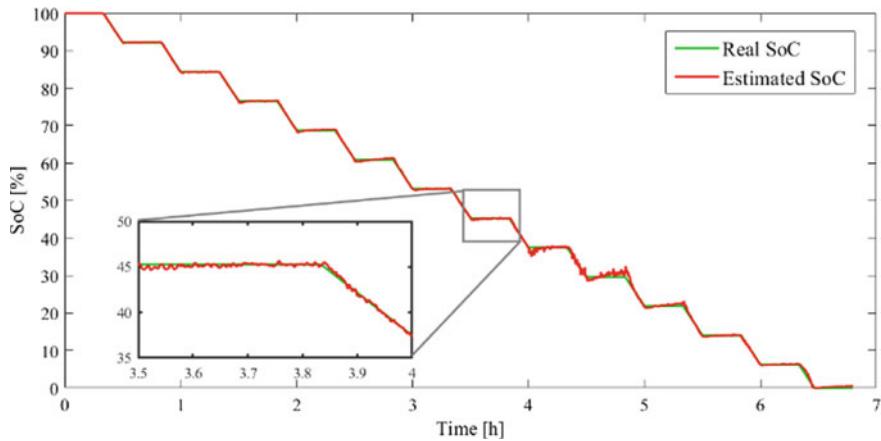
In this section, we provide the experiment of the digital twin model of the battery management system. The performance of the SoC estimation and SoH estimation algorithms are evaluated.

Figure 4 shows the results of SoC estimation result based on Thevenin battery model and OCV method. It is in fact an open-loop estimation method, where the SoC is computed based on the battery mathematical model and OCV. Figure 5 shows the results of SoC evaluation based on the extended Kalman filter approach. Using the Thevenin battery model, the Kalman filter compares the model output with the actual output, which is further used to update the estimation. It can be seen that the SoC estimation method has a good estimation performance.

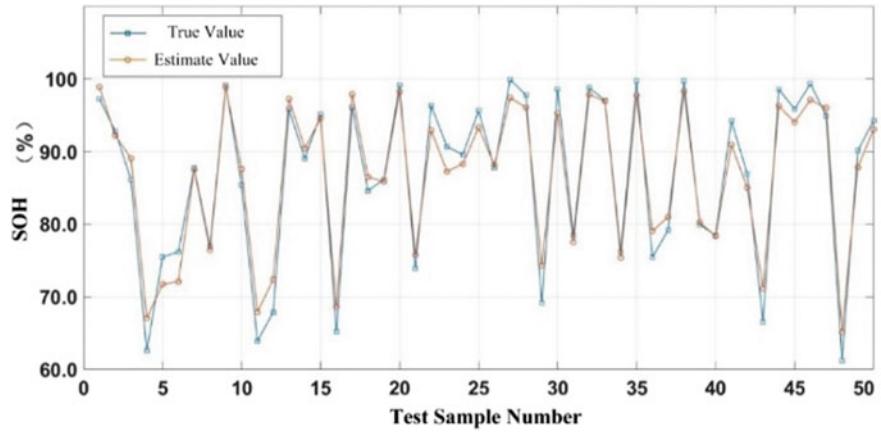
Figure 6 shows the state of health estimation results when the test data of different batteries are intercepted at a shorter voltage segment. The particle optimization



**Fig. 4** The SoC estimation based on Thevenin battery model and OCV method



**Fig. 5** The SoC estimation based on extended Kalman filter



**Fig. 6** The SoH estimation based on the particle optimization method

method is applied to estimate the SoH of batteries. The absolute error of the estimation results of most test samples is less than 5%. It can be seen that the shorter the intercepted voltage segment, the worse the estimation accuracy of the model. Therefore, in order to ensure the estimation accuracy of the model, a longer voltage segment should be selected as much as possible.

## 6 Summary

In this paper, a digital twin model is proposed for battery management systems. We discuss the basic concepts of the digital twin model, including the physical layer and the cyber layer. The digital twin model, algorithms and platforms are presented in detail. The experiment results of the proposed digital twin model show that the SoC and SoH can be estimated accurately. Future work will focus on the development of digital twin model of the battery management system with the cell balancing.

## References

1. Kiehne, H.A.: *Battery Technology Handbook*. CRC Press (2003)
2. Jongerden, M.R., Haverkort, B.R.: Which battery model to use? *IET Softw.* **3**(6), 445–457 (2009)
3. Xiong, R., Li, L., Tian, J.: Towards a smarter battery management system: a critical review on battery state of health monitoring methods. *J. Power Sources* **405**, 18–29 (2018)
4. Trovò, A.: Battery management system for industrial-scale vanadium redox flow batteries: Features and operation. *J. Power Sources* **465**, 228229 (2020)
5. Xiong, R., Ma, S., Li H., et al.: Toward a safer battery management system: A critical review on diagnosis and prognosis of battery short circuit. *Isience* **23**(4), 101010 (2020)
6. Park, J., Lee, M., Kim, G., et al.: Integrated approach based on dual extended Kalman filter and multivariate autoregressive model for predicting battery capacity using health indicator and SOC/SOH. *Energies* **13**(9), 2138 (2020)
7. Cen, Z., Kubiak, P.: Lithium-ion battery SOC/SOH adaptive estimation via simplified single particle model. *Int. J. Energy Res.* **44**(15), 12444–12459 (2020)
8. Xiao, D., Fang, G., Liu, S., et al.: Reduced-coupling coestimation of SOC and SOH for lithium-ion batteries based on convex optimization. *IEEE Trans. Power Electron.* **35**(11), 12332–12346 (2020)
9. Li, W., Rentemeister, M., Badeda, J., et al.: Digital twin for battery systems: Cloud battery management system with online state-of-charge and state-of-health estimation. *J. Energy Storage* **30**, 101557 (2020)
10. Naguib, M., Kollmeyer, P., Emadi, A.: Lithium-ion battery pack robust state of charge estimation, cell inconsistency, and balancing. *IEEE Access* **9**, 50570–50582 (2021)
11. Pham, V.L., Duong, V.T., Choi, W.: A low cost and fast cell-to-cell balancing circuit for lithium-Ion battery strings. *Electronics* **9**(2), 248 (2020)
12. Pham, V.L., Duong, V.T., Choi, W.: High-efficiency active cell-to-cell balancing circuit for Lithium-Ion battery modules using LLC resonant converter. *J. Power Electron.* **20**(4), 1037–1046 (2020)
13. Wei, Z., Zhao, D., He, H., et al.: A noise-tolerant model parameterization method for lithium-ion battery management system. *Appl. Energy* **268**, 114932 (2020)
14. Tao, F., Zhang, H., Liu, A., et al.: Digital twin in industry: state-of-the-art. *IEEE Trans. Ind. Inf.* **15**(4), 2405–2415 (2018)
15. Fuller, A., Fan, Z., Day, C., et al.: Digital twin: enabling technologies, challenges and open research. *IEEE Access* **8**, 108952–108971 (2020)
16. Merkle, L., Segura, A.S., Grummel, J.T., et al.: Architecture of a digital twin for enabling digital services for battery systems. In: 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), pp. 155–160. IEEE (2019)
17. Qu, X., Song, Y., Liu, D., et al.: Lithium-ion battery performance degradation evaluation in dynamic operating conditions based on a digital twin model. *Microelectron. Reliabil.* **114**, 113857 (2020)

18. Bhatti, G., Mohan, H., Singh, R.R.: Towards the future of smart electric vehicles: Digital twin technology. *Renew. Sustain. Energy Rev.* **141**, 110801 (2021)
19. Li, H., Peng, J., He, J., et al.: Synchronized cell-balancing charging of supercapacitors: a consensus-based approach. *IEEE Trans. Ind. Electron.* **65**(10), 8030–8040 (2018)
20. Li, L., Huang, Z., Li, H., et al.: A rapid cell voltage balancing scheme for supercapacitor based energy storage systems for urban rail vehicles. *Electric Power Systems Res.* **142**, 329–340 (2017)
21. Jiang, F., Meng, Z., Li, H., et al.: Consensus-based cell balancing of reconfigurable supercapacitors. *IEEE Trans. Ind. Appl.* **56**(4), 4146–4154 (2020)

# A Research on Remaining Useful Life of Solenoid Valve Based on Millimeter Wave Radar



Xin Liu , Shou Li , Weirong Liu , and Feng Zhou

**Abstract** Since the solenoid valves (SVs) are widely used, it is critical to pay close attention to its operating conditions and remaining useful life (RUL). In this paper, a detection technique for the SV core displacement based on millimeter wave radar is proposed, and the RUL prediction is achieved by auxiliary particle filter (APF) technology. First, the core detection of a single SV is realized by the phase change of the FMCW radar, and the degraded data set is constructed. Second, the exponential model and the Stochastic process model are combined with the APF technology, the model parameters are updated by the latest measurement, and the dynamic model parameters are constructed. Finally, the RUL prediction was completed through the experiment platform. The experiment results verify the reliability of the method proposed herein.

**Keywords** Frequency modulated continuous wave (FMCW) · Auxiliary particle filter (APF) · Solenoid valve (SV) · Remaining useful life (RUL) · Predictable maintenance

## 1 Introduction

The SV is the core component of the electro-hydraulic and electro-pneumatic control system, and its reliability affects the operating efficiency and performance of the entire control system. In order to replace the SV before it fails, to improve the system's reliability and reduce the maintenance cost, the RUL estimation theory is introduced to realize the predictable maintenance of the SV [1].

In recent years, the development trend of RUL prediction is the combination of physical failure model and data-driven fusion model prediction. Xilang et al. [1] defined the SV coil drive current as a degradation indicator, and PF technology was

---

X. Liu · S. Li · F. Zhou

Changsha University of Science & Technology, Changsha 410114, China  
e-mail: [lishuo@csust.edu.cn](mailto:lishuo@csust.edu.cn)

W. Liu  
Central South University, Changsha 410083, China

used to predict its RUL. Yuefeng et al. [2] used battery capacity as a degradation indicator constructed a degradation model of lithium batteries and estimated its RUL combined with PF technique. However, after resampling the particles using the PF technique, the diversity of the particle swarm decreases, resulting in a decrease in the accuracy of the system's prediction. This paper proposes an RUL prediction method based on the APF technique to solve this problem.

According to the working principle of the SV, its core displacement is the most direct reflection of operating conditions, so it is appropriate to select the distortion degree of the core displacement as a degradation indicator. Most of the existing research is to construct the degradation indicator of the SV by indirectly detecting the drive signal. Alexander et al. [3] and Hao et al. [4] used a current sensor to detect the operating current of the SV. They estimated the core displacement through the SV and the inductance model analysis as a degradation indicator to monitor the health of the SV. The current RUL research of SV is mainly based on the indirect measurement of core displacement, which has low measurement accuracy. At the same time, the installation of the detection device will destroy the original structure of the equipment and cannot achieve the non-destructive detection of the SV.

The millimeter-wave FMCW radar has high-sensitivity detection capabilities and can achieve high-precision non-destructive detection of the target's small deformation or high-speed displacement [5]. Sacco et al. [6] and Mostafa et al. [7] proposed a vital sign detection method based on FMCW radar, which extracts vital human signs by detecting small chest cavity displacements caused by heartbeat respiration.

In order to solve the problem of the existing SV RUL prediction, this paper proposes a method based on millimeter-wave FMCW radar and APF technology to realize the lossless RUL prediction of the SV.

## 2 Core Displacement Detection Based on FMCW Radar

The frequency of the chirps transmitted by the FMCW radar is periodic linearly-increasing. Transmitted FMCW signal is

$$T(t) = \exp \left[ j \left( 2\pi f_c t + \frac{\pi B}{T} t^2 \right) \right] \quad (1)$$

Among them,  $f_c$  is the start frequency of the sensor signal,  $T$  is the duration of the chirp, and  $B$  is the bandwidth. The signal at the receiver is a delayed version of the transmitted signal as in

$$R(t) = \exp \left[ j \left( 2\pi f_c (t - \tau) + \frac{\pi B}{T} (t - \tau)^2 \right) \right] \quad (2)$$

The delay time  $\tau = 2R/C$ , Where R is the distance of the target from the radar, and C is the propagation speed of the radar signal. After the transmitted and received signals are mixed and filtered, the beat signal from an object at range R is given by

$$B(t) = T'(t)R(t)\exp\left[j\left(4\pi \frac{BR}{CT}t + \frac{4\pi}{\lambda}R\right)\right] \quad (3)$$

The wavelength  $\lambda = C/f_c$ . If the radar measures the phase change  $\Delta\varphi$  between successive measurements, the target displacement is

$$\Delta d = \frac{\lambda}{4\pi} \Delta\varphi \quad (4)$$

Equation 4 shows that a phase change of 1 rad means that the target is displaced by 0.31 mm, which is sufficient to detect the core displacement.

### 3 APF-Based RUL Prediction

#### 3.1 Definition of Degradation Indicator

The core displacement curve of the SV can directly describe its operation. If the SV performance is degraded or malfunctions, the core's movement will change, which means that the core displacement curve of the SV will be distorted. This paper will use the distortion degree of the SV core displacement as a degradation indicator. In order to describe its degradation indicator, the template core displacement of the SV  $X_{temp}$  at the beginning of operation is defined as follows

$$X_{temp} = [x_{temp1}, x_{temp2}, x_{temp3}, \dots, x_{tempm}] \quad (5)$$

Here  $x_{tempm}$  represents the  $m$ th template displacement waveform sample. When the SV is usually working, the core displacement waveform in the  $k$ th cycle is defined as in

$$X_k = [x_{k1}, x_{k2}, x_{k3}, \dots, x_{km}] \quad (6)$$

Here  $x_{km}$  represents the  $m$ th core displacement waveform sample in the  $k$ th cycle. The Euclidean distance  $d(X_{temp}, X_k)$  between  $X_{temp}$  and  $X_k$  is used to represent the distortion degree of the SV core displacement curve at the  $k$ th cycle.

$$d(X_{temp}, X_k) = \sqrt{\sum_{j=1}^m (X_{kj} - X_{tempj})^2} \quad (7)$$

We use Eq. (7) to calculate the distortion degree of the SV core displacement throughout its life cycle to construct the SV degradation data set.

### 3.2 RUL Estimation of SV Based on APF Technique

The APF technique increases the number of sampled particles by introducing auxiliary variables  $l_k^i$ . It selects the joint probability density function of the auxiliary variable  $l_k^i$  and the measurement  $y_k$  as the importance probability density function and obtains the sampled particles in advance to enhance the credibility of the sampled particles. This paper introduces the exponential model [8] and the Stochastic process model [9] combined with APF technology to carry out RUL prediction experiments.

First, we obtain sampled particles from the prior  $p(\Theta_0)$  to construct the initial particle distribution data set  $\{\Theta_0^i, \omega_0^i\}_N$  ( $\Theta = [x, \theta, v]$ ,  $x$  is the degraded state,  $\theta$  is the model parameter, and  $v$  is the measurement noise). Second, sample from  $p(\Theta_k|\Theta_{k-1}^i)$  to extract auxiliary variables  $l_k^i$ , and get the normalized weights  $\omega_i^A(i)$ . According to the weight  $\omega_i^A(i)$ , the particle set  $\{\Theta_{k-1}^i\}_N$  is resampled to construct a new particle distribution set  $\{\Theta_{k-1}^{j(i)}\}_N$ . Then, the particle  $\Theta_{k-1}^{j(i)}$  is transferred from step<sub>k-1</sub> to step<sub>k</sub> through the state transition equation of the degradation model, and a new set of particles  $\{\Theta_k^i\}$  is generated at step<sub>k</sub>, its normalized weight  $\omega_k^i$  is obtained. Then estimate the degradation state of the SV from the particle state and its weight.

$$\hat{x} = \sum_{i=1}^N \omega_k^i x_k^i \quad (12)$$

Finally, resampling is introduced through the measurement equation. The particles with larger weights are copied, the particles with smaller weights are eliminated, and the degradation model parameters and measurement errors are updated. In this way, invalid sampling particles are reduced, and the estimation performance of the system is improved.

Assuming that the RUL prediction starts at  $t_k$ , we will use the particle state  $x_k^i$  and the model parameter  $\theta_k^i$  to estimate the degradation state of the future particles, and recursively generate future measurement values through the measurement equation of the degradation model. The estimated RUL of the SV at the current moment is defined as

$$t_{end} = \inf\{t_e, x(t_e) >= F\} \quad (13)$$

where  $F$  is the failure threshold,  $t_e$  indicates the moment when the degradation state is greater than the failure threshold, and the function  $\inf\{\cdot\}$  indicates the minimum value of the variable. We denote  $t_{end}$  of the  $i$ th particle as  $t_{end}^i$ . When the degradation states of all particles are greater than the failure threshold, the RUL estimate ends.

We assume that  $RUL_k^i$  represents the prediction of RUL for the  $i$ th particle at  $t_k$ , then  $RUL_k^i$  can be expressed as

$$RUL_k^i = t_{end}^i - t_k \quad (14)$$

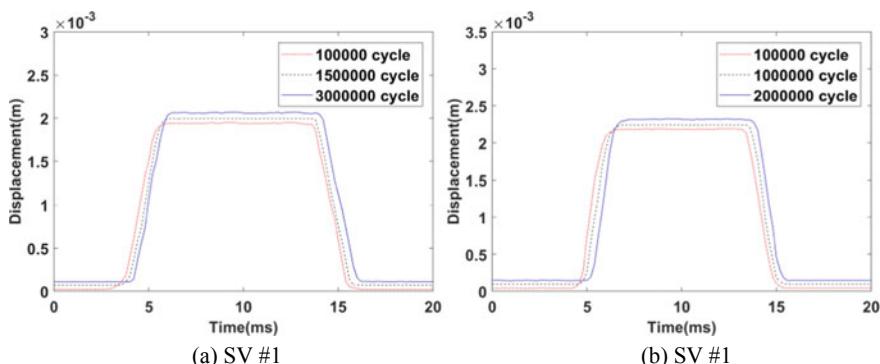
Through Eq. (14) we can get the RUL distribution of the SV at  $t_k$ .

## 4 Experiment and Analysis

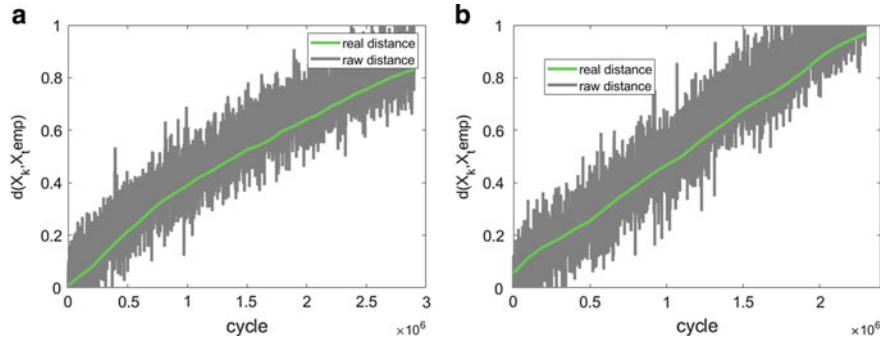
Two SVs (SV#1 and SV#2) with a cycling frequency of 50 Hz were used for the degradation experiment to evaluate the proposed method's feasibility. Table 1 shows the FMCW radar parameters used in the experiment.

**Table 1** FMCW radar parameters

Parameters	Value
Starting frequency $f_c$ (GHz)	77
Bandwidth B (GHz)	4
Number of TX antennas used	1
Number of RX antennas used	1
TX antenna gain (dB)	8
RX antenna gain (dB)	8
Fast time axis sampling	2 MHz
Slow time axis sampling	20 Hz
Chirp duration T	50 us
ADC samples per chirp	100



**Fig. 1** Core displacement curve difference versus cycle number



**Fig. 2** Degradation indicator throughout the life cycle

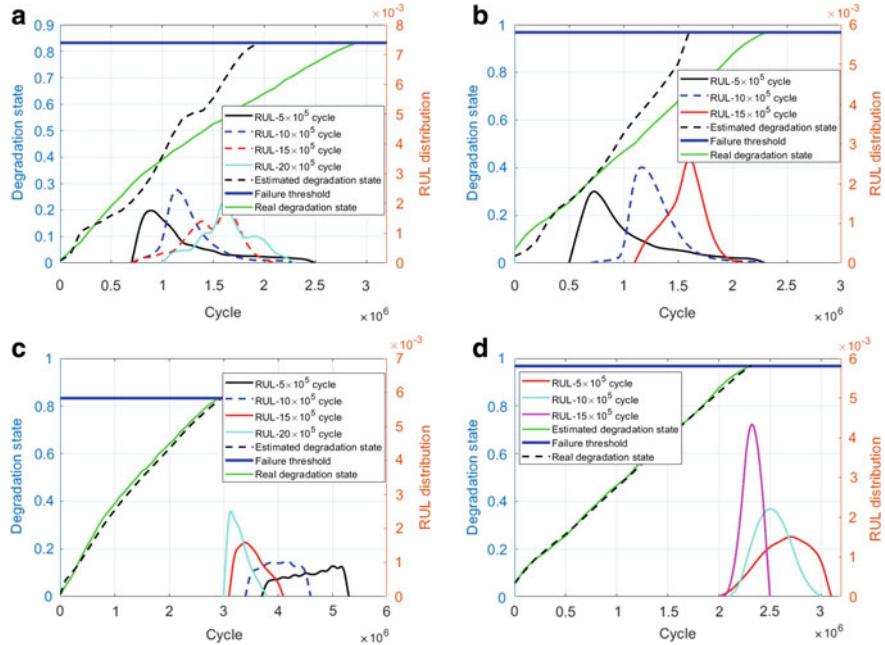
Figure 1 shows the core displacement curves of SV#1 and SV#2 in different cycles. As the working cycle increases, the core displacement curve will change slowly. The core displacement gradually increases, resulted from the aging of the spring and the wear of the valve cavity. At the same time, the response time of the SV is getting longer. The Euclidean distance  $d(X_{temp}, X_k)$  calculates by Eq. (7) is plotted as the raw distance curve in Fig. 2. The noise is too strong to reflect the degradation trends of the SV. In order to obtain its degradation trend, smoothing and filtering processes are used, represented by the real distance in Fig. 2.

In this experiment, the real distance is used as the SV degradation data set, and the experimental parameters are set as follows. The initial degradation state  $x_0$  of SV#1 and SV#2 are 0.0086 and 0.056, respectively. The failure threshold is set to 0.837 and 0.968, respectively. The particle number N is set to be 5000. The SV#1 and SV#2 degeneration process runs to failure at 3,063,120 and 2,236,540 cycles, respectively. The initial parameters of the degradation model are obtained based on the degradation data set and the maximum likelihood estimation algorithm, and the model parameters are updated in real-time through APF technology to construct the dynamic model parameters.

Table 2 shows the RUL prediction results using the exponential model and the Stochastic process model, respectively. And their estimated degradation process and RUL distribution are shown in Fig. 3. From Table 2, we can see that the RUL estimate

**Table 2** RUL prediction results

	Start cycle	Exponential model	Stochastic process model	Real RUL
SV #1	500,000	759,940	3,855,180	2,563,120
	1,000,000	546,790	2,725,410	2,063,120
	1,500,000	349,700	1,912,790	1,563,120
	2,000,000	Failure	1,267,790	1,063,120
SV #2	500,000	601,160	1,869,810	1,736,540
	1,000,000	395,410	1,425,750	1,236,540
	1,500,000	142,510	862,800	7,365,040



**Fig. 3** State estimation and RUL distribution

based on the exponential model is far from the real RUL. In contrast, the RUL estimation based on the stochastic model process model is closer to the real RUL. We can also get the same conclusion from the estimated degradation state in Fig. 3.

It can be seen from the above that it is unreasonable to use the exponential model to predict the degradation process and its RUL, while the stochastic process model is excellent in this respect. However, the RUL prediction performance based on the stochastic process model is not satisfactory in the early stage, which can be seen from the RUL prediction value in Table 2 and the RUL distribution in Fig. 3. As the starting prediction time shifts, the estimated RUL will be closer to the real RUL, and the RUL distribution will become more concentrated. This phenomenon is because as the available measurements increase, the model parameters estimated based on the APF will be more in line with the real degradation. The experimental results show that the SV RUL prediction based on the stochastic process model and APF performs well.

## 5 Conclusions

In this paper, the FMCW radar is used to measure the SV core displacement in real-time, and the distortion degree of the core displacement curve is used as a

degradation indicator to reflect the performance of the SV. Two degradation models combined with APF Technique were used to conduct comparative experiments on RUL prediction. The results show that the FMCW radar can accurately detect the SV core displacement in real-time. The degradation experiment based on the stochastic process model and APF technique can reflect the real degradation state of the SV. This method can calculate accurate RUL and provide a basis for real-time monitoring of SV's reliability to ensure the stable operation of the control system.

## References

1. Xilang, T., Mingqing, X., et al.: Application of particle filter technique to online prognostics for solenoid valve. *J. Int. Fuzzy. Syst.* **35**(1), 1–10 (2018)
2. Yuefeng, L., Guangquan, Z., et al.: A fusion prediction method of lithium-ion battery cycle-life. *Chin. J. Sci. Instrum.* **36**(07), 1462–1469 (2015)
3. Alexander C.Y., James D.C.: Predicting solenoid valve spool displacement through current analysis. *Int. J. Fluid Power* **16**(3):133–140 (2015)
4. Hao, T., Yuren, Z.: Coil inductance model based solenoid on-off valve spool displacement sensing via laser calibration. *Sensors* **18**(12), 4492–4505 (2018)
5. Zhenyu, L., Huiming, C., Wei, L., et al.: Radar vital signal detection based on improved complete ensemble empirical mode decomposition with adaptive noise. *Chin. J. Sci. Instrum.* **039**(012), 171–178 (2018)
6. Giulia, C., Emanuele, P., et al.: An FMCW radar for localization and vital signs measurement for different chest orientations. *Sensors* **20**(12), 3489–3502 (2020)
7. Mostafa, A., George, S., et al.: Remote monitoring of human vital signs using mm-wave FMCW radar. *IEEE Access* **7**, 54958–54968 (2019)
8. Fazludeen, R., Sundararajan, A.: Application of particle filter to on-board life estimation of LED lights. *IEEE Photon. J.* **9**(3), 1–16 (2017)
9. Yaguo, L., Naipeng, L., Jing, L.: A new method based on stochastic process models for machine remaining useful life prediction. *IEEE Trans. Instrum. Mesurem.* **65**(12), 2671–2684 (2016)

# A Multi-link Data Congestion Control Algorithm in Spatial Delay Tolerance Network



Li Yi and Renjie Zhang

**Abstract** The overall control effect of the current DTN data congestion control-related results and the data transmission success rate performance are to be optimized. It is proposed to introduce the ant colony algorithm into the multi-link data congestion control in the spatial delay tolerant network. Based on the overall state of the data cache, the nodes in the spatial delay tolerant network are divided into unsaturated and saturated nodes. If the node's remaining cache cannot hold a single piece of data, then the node belongs to a saturated node; instead, it is a non-saturated node. The state of each node of the communication range is obtained by using the judgment result, and the data congestion in a given area is perceived in real time. Based on network state awareness, the ant colony algorithm is introduced to form a spatial delay tolerant network routing scheme by using the forwarding and copying data distribution method. When the number of replicas is greater than 1, the replication scheme is used, and the forwarding scheme is used to determine whether to forward the data packet or not. Solutions to improve data congestion control performance. Considering the situation of frequent and intermittent connections in the network of spatial delay tolerance, a corresponding route maintenance strategy is proposed. The experimental results show that the proposed algorithm can control DTN data congestion well, and the data transmission success rate is high and reliable.

**Keywords** Information processing · Spatial delay · Many links · Congestion control

## 1 Introduction

DTN (delay tolerant network) is a new network structure, which has the characteristics of high delay and low data rate. In order to realize the interconnection between heterogeneous networks and the stable transmission of asynchronous data, DTN provides a series of mechanisms such as store and forward [1, 2]. The network has a

---

L. Yi · R. Zhang

Hunan Post and Telecommunication College Network Center, Changsha 410015, China

e-mail: [Liyi8984e@126.com](mailto:Liyi8984e@126.com)

very wide range of application prospects, such as sensor networks and global mobile networks, and has been concerned and deeply studied by academia and industry. Due to the great difference between DTN and DTN, the former routing and congestion control schemes are not suitable for DTN [3]. In conclusion, many scholars have been involved in the research of delay tolerant network link data congestion control.

Shi and others proposed a network multi-link data congestion control algorithm based on QoS [4]. In the process, the algorithm is divided into contact congestion judgment and data forwarding under QoS. According to the remaining available contact capacity and the remaining data storage space of nodes, the data congestion of each contact is predicted, and the contacts are divided into different congestion levels. In the process of routing calculation, the highest congestion level of data in a whole path is determined as the congestion level of the path, and traffic data with different priorities are transmitted based on the congestion level. Tan Jing proposed a delay tolerant network infection routing method with congestion control scheme. In the process, according to the dynamic storage state model, the threshold of semi congestion is adjusted to alleviate the data congestion [5]. Add the ACK index and message control queue, promote the node storage to update with the overall load of the network, and delete redundant packets. Based on the advantages of infection routing and prophet routing in different congestion situations, select single or mixed form to realize message forwarding, so as to prevent and relieve data congestion, and complete data storage control and network congestion control. Cai Yueping and others proposed using dftcp to control network congestion [6]. According to the active queue management mechanism, dftcp transmits congestion information through explicit congestion notification methods and uses the queue length of the control switch to alleviate the packet loss of burst traffic, so as to solve the TCP incast problem. In addition, DFTCP clusters TCP flows in terminal service equipment. Once the network congestion occurs, it uses network state information and traffic classification information to adjust the TCP congestion window correspondingly and uses congestion backoff strategy to reduce the impact on other TCP flows in the network, so as to reduce the delay.

The research results of data congestion control in delay tolerant networks need to be optimized in terms of congestion situation and data transmission success rate. A multi-link data congestion control based on Ant Colony Algorithm in spatial Delay Tolerant Networks is proposed.

## 2 Multi-link Data Congestion Control in Spatial Delay Tolerant Networks

Firstly, real-time sensing of data congestion in a given area is completed to enhance the performance of multi-link data congestion control, reduce the congestion rate and improve the success rate of data transmission. In this paper, the ant colony algorithm is used to improve the performance of data congestion control.

## 2.1 A Subsection Sample Node State Awareness in Spatial Delay Tolerant Networks

Spatial delay tolerant network is a research hotspot in the field of wireless network in recent years. Figure 1 shows a typical application scenario of spatial delay tolerant network.

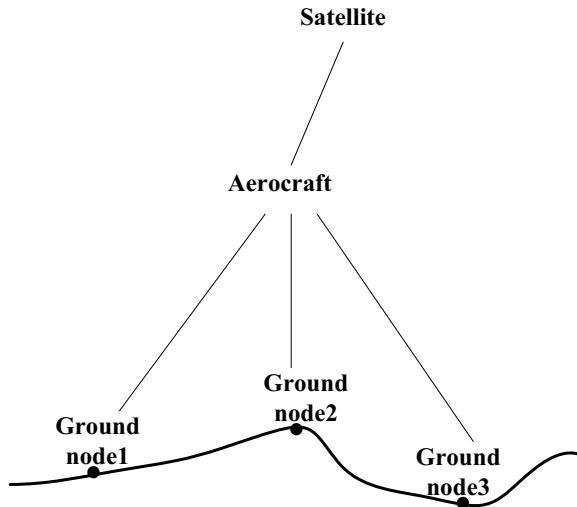
Here, the DTN multi-link data congestion is controlled. In fact, the overall occupancy of node cache is directly related to multi-link data congestion. However, for spatial Delay Tolerant Networks, single node congestion can not fully indicate that other nodes in its subordinate region also have congestion [7, 8]. Therefore, the following steps are used to sense the state of the network nodes, and the whole network situation is perceived in a non-contact way, which lays the foundation for multi-link data congestion control.

Based on the overall state of data cache occupancy, nodes in space delay tolerant networks can be divided into unsaturated and saturated nodes. If the node's remaining cache can't hold a single data, the node belongs to saturation node; On the contrary, it is unsaturated node. Using the state of each node in the communication range, we can sense the data congestion in a given area in real time.

In an ideal case, the upper limit of the number of saturated nodes transformed from unsaturated nodes of all meeting nodes in a given time period  $N_{\max}(t)$  is the total number of unsaturated nodes they meet in  $[t, t + \Delta t]$ , which can be expressed in the form of Eq. (1):

$$N_{\max}(t) = N_c(t) \quad (1)$$

**Fig. 1** Typical spatial delay tolerance network application scenario



where  $N_c(t)$  represents the total number of the encounter nodes in the unsaturated state at  $t$ , and there is  $N_c(t) = N_t(t) - N_s(t)$ . Where  $N_t(t)$  represents the number of encounter nodes at  $t$  time, and  $N_s(t)$  represents the number of saturation nodes of all encounter nodes at  $t$  time. Assuming that  $N_t(t)$  is 0, it means that the node does not create a connection with other nodes. If the current maximum connection duration is  $\Delta t$ , then the maximum data flow of this connection can be expressed as  $B \times \Delta t$ , where  $B$  represents the channel transmission rate. Assuming that the node's remaining cache is larger than  $B \times \Delta t$ , the node's state transition probability is 0; On the contrary, if a node has a state transition, the upper limit of the amount of received data  $M_{\max}$  can be expressed as:

$$M_{\max} = \frac{B \times \Delta t}{m_{size}} \quad (2)$$

where  $m_{size}$  represents the data size.

To sum up, based on the basic principle of spatial delay tolerant network data forwarding, it can be observed that after any node  $n_k$  and  $n_q$  meet, if there are at least  $M_{\max}$  of  $n_k$  in node  $n_q$  that have not been saved, the state of  $n_k$  will change. Thus, the  $n_k$  state transition probability can be expressed as:

$$P_r(n, q) = \begin{cases} \frac{n}{q} \times \frac{M_{\max}^{n_k}}{M_{\max}^{n_q}}, & M_{\max} \leq n < q \\ 0, & 0 \leq n < M_{\max} \end{cases} \quad (3)$$

where,  $q$  represents the amount of data carried by  $n_q$ .

To sum up, the actual number of state transitions in unsaturated nodes can be expressed as:

$$N_{act} = P_r(n, q) \times N_{\max}(t) \quad (4)$$

According to the above ideal situation, because the saturated node cache cannot hold additional copies of data, its overall state will not change. However, in the specified time range of  $\Delta t$ , due to the node actively discarding the data in the local cache at the end of the active time, some saturated nodes in the network will be converted to unsaturated nodes. To sum up, the number of saturated nodes in the next moment mainly depends on two aspects: one is that the unsaturated nodes are transformed into saturated nodes after receiving the data, and the other is that the saturated nodes are transformed into unsaturated nodes after deleting the active expiration data. Therefore, the number of saturated nodes in the process of node encounter at the next moment can be expressed as:

$$N_s(t + \Delta t) = N_{act} + (N_s(t) - N_s(t) \times P(t_{\min} < \Delta t)) \quad (5)$$

In the formula,  $N_s$  represents the number of saturated nodes of all encounter nodes, and  $P(t_{\min} < \Delta t)$  represents the probability value of data discarded by saturated nodes in the time range.

Related studies show that in a relatively short period of time, the number of nodes that can meet a given node follows Poisson distribution, that is to say, the number of nodes that can meet a given node follows a generalized stationary random process [9, 10].

According to the above calculation and analysis, we can estimate the number of encounter nodes at the next moment through historical information, and predict the number of encounter nodes according to the exponential smoothing method in the generalized stationary random process. Here,  $\alpha$  is defined as the weight, that is, the probability value of nodes meeting, which is used to estimate the number of nodes meeting at the next moment:

$$S_{t+1}^{di} = |\alpha \times N_t(t) + (1 - \alpha) \times S_{t-1}| \quad (6)$$

where  $S_{t-1}$  represents the predicted value of the previous moment. Therefore, the probability of nodes meeting in  $\Delta t$  can be expressed as follows:

$$\alpha = 1 - e^{-\Delta t} \quad (7)$$

According to the above calculation, it can be clearly observed that there is a boundary in the range of motion of the network nodes within the specified time  $\Delta t$ . Therefore, it can be determined that the overall topology of the delay tolerant network in the area where the nodes are located is relatively stable. The formula for calculating the number of unsaturated nodes in the next moment is as follows:

$$N_v(t + \Delta t) = S_{t+1}^{di} - N_s(t + \Delta t) \quad (8)$$

According to Eq. (8), the number of unsaturated nodes in the network is calculated, the status of each node in the communication range is obtained, and the real-time sensing of data congestion in a given area is completed.

## 2.2 Data Allocation and Routing Maintenance Based on Ant Colony Algorithm

According to 2.1 network node perception, this paper will use ant colony algorithm to achieve multi-link data congestion control. The advantages of this algorithm are that the method is simple, easy to implement, and the search value is good. The specific implementation process is as follows: set the perception range, detect the environmental information, find food according to the ant colony's foraging rules, namely pheromone, track the characteristics of ants moving towards pheromone

and obstacle avoidance rules, and find the optimal foraging path. multi-link data congestion control in spatial Delay Tolerant Networks via data allocation and routing maintenance.

### 2.2.1 Data distribution

Data forwarding scheme.

Set the network nodes  $v_i$  and  $v_j$  to meet at  $t$  time. Considering the random packet  $m'$  in  $v_i$  queue, the main problem is whether the  $v_i$  forwards to  $v_j$  or not. For the target node  $v_k = d(m')$ ,  $\tau_{i,j}^k(t) > \tau_{j,i}^k(t)$  represents contact  $(v_i, v_j)$ , and the overall pheromone content is higher than contact  $(v_j, v_i)$ , that is, the possibility of using  $(v_i, v_j)$  to reach the target node at this moment is higher than  $(v_j, v_i)$ . However, similar to ant foraging, if ants encounter a fork in the process of foraging, they will choose the path according to the pheromone content. For the purpose of reflecting this feature, that is, when  $\tau_{i,j}^k(t) \leq \tau_{j,i}^k(t)$  and  $v_i$  determine whether to forward the packet in the form of probability. At the same time, the higher the content of path pheromone, the higher the probability of being selected. To sum up, for the target node  $v_k = d(m')$ ,  $p_{i,j}^k(t)$  is defined here to represent the probability that  $v_i$  will forward the random packet  $m'$  to  $v_j$ . The expression is:

$$p_{i,j}^k(t) = \begin{cases} \frac{\tau_{i,j}^k(t)}{\sum\limits_{v_j \in N'_i} \tau_{i,j}^k(t)}, & v_j \in N'_i \\ 0, & v_j \notin N'_i \end{cases} \quad (9)$$

where  $N'_i$  represents the set of  $v_i$  hop neighbors. Here,  $f_{i,j} = \{0, 1\}$  is set to represent whether  $v_i$  forwards packets to  $v_j$  or not, and  $\beta$  is defined to represent the random number generated by  $v_i$ . To sum up, the forwarding scheme for multi-link data congestion in spatial Delay Tolerant Networks can be expressed as follows:

$$f_{i,j} = \begin{cases} 1, & \tau_{i,j}^k(t) > \tau_{j,i}^k(t) \text{ or } \beta \leq p_{i,j}^k(t) \\ 0, & \text{else} \end{cases} \quad (10)$$

In addition to Eq. (10), if node  $v_i$  meets with multiple neighbor nodes at the same time, it is judged one by one according to the descending order of forwarding probability.

Replication scheme.

Because the network resources are limited, so the routing algorithm is also run in the case of quota. Based on this,  $v_i$  and  $v_j$  meet at  $t$  time. For random packet  $m'$  in  $v_i$  queue, the main problem is how many copies  $v_i$  should copy and forward to  $v_j$ . Set  $v_i$  to forward the replication packet to  $v_j$  only when there is no copy of the packet in  $v_j$ . Set the number of copies of  $m'$  in  $v_i$  to  $L$ , which is greater than 1, then the  $c_{i,j}$  expression of the number of copies required is:

$$c_{i,j} = \left\lfloor L \times \frac{1}{(\delta + 1)} \right\rfloor \quad (11)$$

Among them,  $\delta$  represents the sum of pheromones. In summary, the number of remaining copies of  $v_i$  is  $L - c_{i,j}$ . Assume that  $L$  is 1, In other words,  $v_i$  only needs to determine whether to forward the packet or not, and then it helps to realize the routing decision according to the forwarding scheme.

According to the above calculation and analysis, this paper introduces the routing schemes in forwarding and replication modes respectively.  $f_{i,j}$  and  $c_{i,j}$  represent the output results of the two modes, that is, when the number of replicas  $L$  is greater than 1, the replication scheme is used. On the contrary, the forwarding scheme is used to determine whether to forward the packets, and the dual control scheme is used to further improve the performance of data congestion control.

### 2.2.2 Route maintenance

Due to frequent breaks and intermittent connections in spatial Delay Tolerant Networks, each node mainly updates all kinds of information in the routing table to adapt to the changing topology. Using the pheromone dissemination and volatilization rules in ant colony algorithm, the pheromone content related to contact can be dynamically adjusted to provide support for routing maintenance.

In the pheromone dissemination rules above, ants are supposed to spread the most pheromones when they find food or nest, and the number of pheromones will decrease with the increasing distance. Therefore, the corresponding dissemination pheromone of  $m'$  will change after one transmission. The details can be expressed as follows:

$$\Delta m' = \alpha \times m', \alpha \in (0, 1) \quad (12)$$

In the pheromone volatilization rule, the pheromone content related to connection will volatilize gradually with the extension of time. If a certain connection is interrupted for a long time, the corresponding pheromone amount will continue to decrease until it is invalid. Pheromone volatilization can be expressed as the following model.

$$\tau(t+1) = (1 - p) \times \tau(t), p \in (0, 1] \quad (13)$$

The parameters in Eq. (13) will be adjusted adaptively with the change of network scale and topology. In order to verify the effectiveness and feasibility of the proposed method, simulation experiments are needed.

### 3 Experimental Results and Analysis

In order to verify the overall performance of multi-link data congestion control in spatial delay tolerant network based on ant colony algorithm, a correlation test is carried out. During the experiment, NS2 network is used to build the experimental platform, and the performance of the proposed algorithm is tested in different data cache storage space and bandwidth simulation environment.

The parameters used in the test are as follows:

1. The model of the experimental node is two way ground;
2. The transmission rate of network channel is 2 Mbit/s;
3. The network routing protocol is epidemic;
4. The effective transmission distance is defined as 50 m;
5. The listening distance is defined as 250 m;
6. The test time is 2600 s;
7. The network coverage is defined as  $1000 \text{ m} \times 500 \text{ m}$ ;
8. The number of mobile nodes is set to 60, and the maximum speed of each node is set to 20 m/;
9. The number of experiments was set to 35.

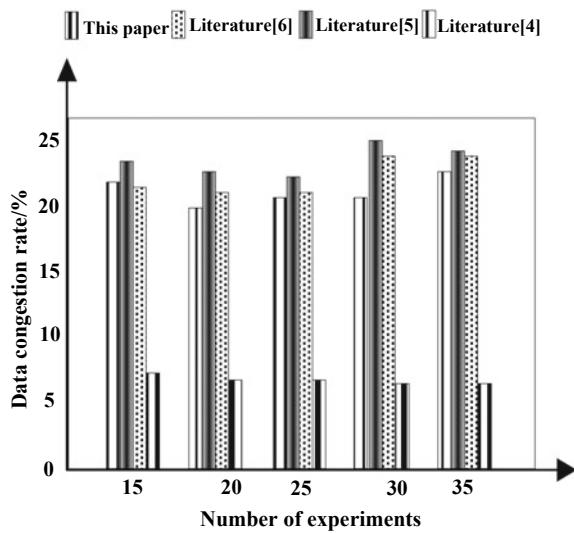
Based on the above experimental parameters, 50 out of 60 nodes are randomly selected as data sources, and the experimental indicators are network data congestion rate and data transmission success rate. The calculation formula of network data congestion rate is as follows:

$$L_N = \frac{N_v(t + \Delta t) P_r(n, q)}{\Delta t} \quad (14)$$

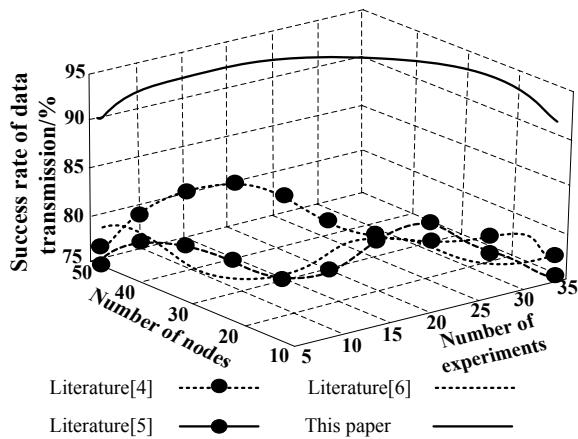
$$P_m = \frac{c_{i,j} \Delta m'}{\Delta t} \quad (15)$$

By analyzing the experimental results in Figs. 2 and 3, compared with the literature results, the data congestion rate under multi-link data congestion control in spatial delay tolerant network based on ant colony algorithm is lower, and the data transmission success rate is higher, and the performance is superior. By sensing the state of the network nodes, the algorithm can perceive the overall situation of the network in a non-contact way, initially enhance the performance of data congestion control, and lay the foundation for reducing the congestion rate and improving the success rate of data transmission. Based on the network state perception, the ant colony algorithm is introduced to control the data congestion by using the routing scheme under the forwarding and replication modes. Through the dual control scheme, the performance of data congestion control is further improved. At the same time, the routing maintenance strategy is used to adjust adaptively with the change of network size and topology, so as to better guarantee the smoothness of multi-link data transmission.

**Fig. 2** Comparison of congestion control of different results network data



**Fig. 3** Comparison of data transmission success rate of different research results



## 4 Conclusion

Due to the data congestion in the process of network communication, which affects the normal operation of the data, it is urgent to control the phenomenon, and which control method is the focus of current research. In this paper, ant colony algorithm is proposed to control multi-link data congestion in space delay tolerant network. The node state perception of spatial delay tolerant network determines the stability of the overall topology of the network, and calculates the number of unsaturated nodes in the next moment. Considering the frequent fracture and intermittent connection in the network, ant colony algorithm is used for data allocation and routing maintenance,

and finally, network congestion control is realized. The experimental results show that the proposed algorithm has strong control performance and robustness. In the next step, we can analyze the data packet loss and throughput in the transmission process to better meet the demand of congestion control in space delay tolerant network.

## References

1. Nunes, I.O., Celes, C., Nunes, I., Vaz de Melo, P.O.S., Loureiro, A.A.F.: Combining spatial and social awareness in D2D opportunistic routing. In: IEEE Communications Magazine, vol. 56, no. 1, pp. 128–135 (2018)
2. Spaho, E.: Energy consumption analysis of different routing protocols in a delay tolerant network. J. Ambient Intell. Hum. Comput. (2019) (prepublish)
3. Nag, S., Li, A.S., Ravindra, V., et al.: Autonomous scheduling of agile spacecraft constellations with delay tolerant networking for reactive imaging. [arXiv:2010.09940](https://arxiv.org/abs/2010.09940) (2020)
4. Shi, W., Gao, D., Zhou, H.: QoS based congestion control for space delay/disruption tolerant networks. J. Electron.: Inf. Technol. **38**, 2982–2986 (2016)
5. Tan, J., Dong, C., Wang, H.: ERC2: DTN epidemic routing method with congestion control strategy. J. Comput. Appl. **39**(1), 32–38 (2019)
6. Cai, Y., Zhang, W., Luo, S.: Differentiated flow transmission control protocol in data center networks. J. Xi'an Jiaotong Univ. **51**(6), 122–128, 152 (2017)
7. Pushparaj, J., Soumya, J.: A link fault tolerant routing algorithm for mesh of tree based network-on-chips, In: 2019 IEEE international symposium on smart electronic systems (iSES) (formerly iNiS), pp. 181–184 (2019)
8. Chen, Y., Duan, Z.: Congested link inference algorithms in dynamic routing IP Network. Math. Prob. Eng. ( 2017)
9. Xianyou, Z., Ying, Y.: Sensor network delay tolerant routing algorithm based on Markov. Bull. Sci. Technol. (2017)
10. Hua, J., Ge, X., Zhong, S.: FOUM: a flow-ordered consistent update mechanism for software-defined networking in adversarial settings. In: IEEE INFOCOM 2016—The 35th Annual IEEE International Conference on Computer Communications. IEEE (2016)

# Data-Driven Fault Prognosis for Pneumatic Valves in Train Electropneumatic Brake System



Dianzhu Gao , Jun Peng , Ning Ding , and Yingze Yang

**Abstract** Pneumatic valves are key components of the train electro-pneumatic braking system. In order to obtain health indicators of pneumatic valves and provide faults early-warning, this paper proposes a fault prognosis method using principal component analysis (PCA) and support vector regression (SVR). Two health indicators ( $T^2$  and SPE) of pneumatic valves are extracted through PCA method based on the full life cycle data set, which came from the joint simulation model. Second, a pneumatic valve fault prognosis model based on SVR is trained based on the health indicators. Combined with the working model of the train electro-pneumatic braking system, the proposed fault prognosis model can estimate the expected time of pneumatic valve fault time accurately. Results from a semi-physical simulation verification platform of DK-2 braking system indicate that the proposed method can effectively predict the occurrence of faults. This work can provide a scientific basis for the operation of braking system and maintenance strategy of pneumatic valves.

**Keywords** Pneumatic valve · Principal component analysis · Support vector regression · Fault prognosis

## 1 Introduction

As the key link of circuit-pneumatic-mechanical control loop in the train braking system, pneumatic valve plays an important role in system's safety and reliability [1]. Pneumatic valves gradually deteriorate with the accumulation of working time, thus leading to a decline in performance [2]. Due to the complex structure, the internal conditions of pneumatic valves cannot be directly observed, resulting in difficulty in predicting when the faults occur.

---

D. Gao · N. Ding

School of Automation, Central South University, Changsha 410075, China

J. Peng · Y. Yang

School of Computer Science and Engineering, Central South University, Changsha 410075, China  
e-mail: [yangyingze@csu.edu.cn](mailto:yangyingze@csu.edu.cn)

Current fault prediction methods consist of physical model-based methods, statistical model-based methods and machine learning-based methods [3]. The physical model-based methods use the thresholds generated by the actual and model values as indicators for prediction [4]. Chetan et al. [5] used a mathematical model describing the physical principles of component degradation, a Bayesian filter algorithm was used for parameter state estimation and fault prediction. The physical model-based methods require accurate mathematical models, but it is difficult to build an accurate mathematical model for pneumatic valves [6].

Instead of building a physical model, the statistical methods use historical data to establish health indicators of the degradation process [7]. Jin et al. [8] derived an auto-regressive model for filtering fault-independent signals to track the degradation process of the bearing and designed an extended Kalman filter for fault prediction. Principal component analysis is widely used in the field of fault prediction where the degradation state can be accurately measured according to the degree of deviation of the statistical indicators [9].

In recent years, machine learning-based methods for fault prediction have drawn a lot of attention, since only enough historical fault data is required to implement complex fault prediction problems [10]. For example, long and short-term memory recurrent neural networks are used to perform multi-forward voltage prediction for battery system failures prediction [11]. Hack-Eun et al. [12] introduced a support vector machine classifier to estimate the health state and to make long-term predictions of the bearing degradation state.

The purpose of this paper is to predict the degradation process in pneumatic valves and to provide early warning before a fault occurs. A support vector regression-based fault prediction method for pneumatic valves is proposed, and different fault prediction models are established for different types of faults. Firstly, the similarities in the failure characteristics of different pneumatic valves are discussed, and the features are reconstructed according to the fault types. Then a principal component analysis method is used to extract the health indicators. Finally, the support vector regression model is trained separately for each fault type to predict the time of fault occurrence.

The remainder of this paper is organized as follows. The degradation and simulation models of pneumatic valves are established in Section II. Section III describes the SVR-based fault prognosis method. Simulation results are presented in Section IV. We conclude the paper in Section V.

## 2 Modeling

In this section, the structure and working principle of the pneumatic valve are analyzed. The simulation models of the train/equalization control unit of the electro-air braking system based on AMESim and Matlab are built to analyze the different fault types and performance, and obtain the fault data to verify the proposed method.

Pneumatic valve is a mechanical valve, which does not rely on electromagnetic drive. In the braking system, pneumatic valves mainly include relay valves, and other valves. Relay valve, as one of the core parts of braking system, is very representative in different types of pneumatic valves. Therefore, this paper takes the relay valve as the research object to verify the effectiveness of the method. The brake control unit is built with Simulink toolbox in MATLAB.

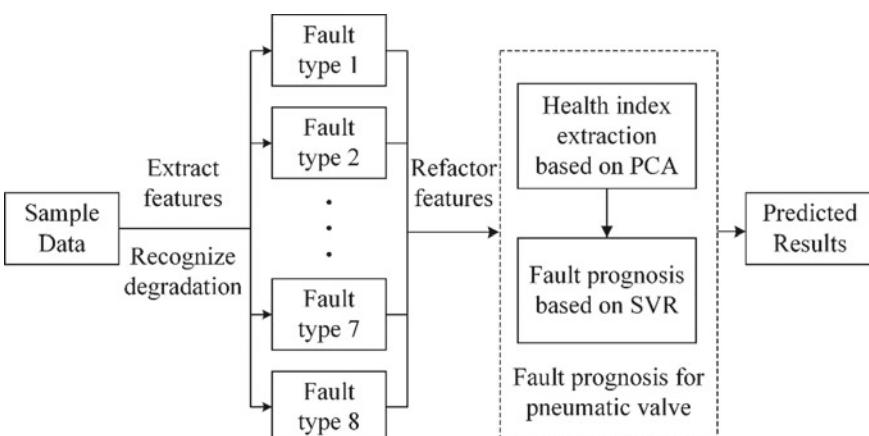
The simulation model of train/equalization control unit is realized by AMESim and Simulink toolbox in MATLAB. The model includes two parts: air channel and brake control unit. The air channel is simulated by the Pnuematic Kit and the Pnuematic Component Design Kit in AMESim.

The AMESim simulation model of the relay valve consists of three parts: the air supply valve, the air evacuation valve and the main piston. The function of the relay valve is to control the pressure of the train pipe according to the pressure of the equalizing reservoir, and to realize the functions of air inflation, exhaust and maintaining pressure.

The inflation function is to increase the pressure in the train tube so that the train can run normally. The exhaust function is to reduce the pressure in the train tube to slow down or stop the train. The pressure maintaining function is to keep the train in the braking state.

As shown in Fig. 1, this paper proposes a fault prognosis method for pneumatic valve based on health indicators and support vector regression. Different fault prognosis models are trained with the data generated by different fault types. The importance of the degradation features was evaluated, and a group of features with the highest weight for each fault type was selected as sample data. Principal component analysis is used to calculate statistics  $T^2$  and SPE as health indicators. The prognosis model is established based on the support vector regression algorithm.

It mainly consists of two parts, features extraction using principal component analysis and model establishment based on support vector regression.



**Fig. 1** The proposed fault prognosis framework for pneumatic valve

### 3 Fault Prognosis Using Health Index Extraction and Support Vector Regression

Pneumatic valves are key components of train braking system. Effective fault prediction method of pneumatic valves can avoid the failure of pneumatic valve and ensure the safety and reliability of train operation. Therefore, a fault prediction model of pneumatic valve based on health index and support vector regression is proposed in this paper.

#### 3.1 *Health Index Extraction Based on Principal Component Analysis*

After the sample data reconstruction based on the importance of fault features, the health indicators corresponding to the data samples should be extracted according to the fault symptom types of the samples. Aiming at the problem of pneumatic valve health indicators are difficult to extract, this section on the basis of the characteristics of the reconstruction, this paper proposes a pneumatic valve health indicators on the basis of the principal component analysis model, mainly through the orthogonal transformation, the original data space decomposition is given priority to yuan space and residual space, while reducing the data dimension reserves the main information, and then extracted respectively from two Spaces health indicators.

The reconstructed degradation characteristics of pneumatic valves were used as the input data of the health index extraction method. Given a two-dimensional data matrix  $X \in \mathbb{R}^{n \times m}$ ,  $n$  is the number of observations,  $m$  is the number of process variables included in the principal component analysis.

$$X = \hat{X} + E = TP + E = \sum_{i=1}^m t_i p_i^T + E \quad (1)$$

where  $\hat{X}$  is the cross product sum of subspace,  $t_i$  is the principal component vector;  $p_i$  is the projection direction of the principal component by transforming the basis vector;  $E \in \mathbb{R}^{n \times m}$  is the residual matrix of the prediction error of the principal component model.  $E$  is measures the weak trend in the change of uncertainty and degradation process. According to the change trend, two statistics ( $T^2$  and SPE) are obtained to detect the system state, as shown in Eqs. (2) and (3):

$$T_k^2 = \sum_{i=1}^A \frac{t_{ki}^2}{\sigma_i} \quad (2)$$

$$Q_k = \sum_{j=A+1}^m t_{kj}^2 \quad (3)$$

where  $T_k^2$  and  $Q_k$  are the sample of  $T^2$  and SPE at time  $k$ ;  $t_{ki}^2$  and  $t_{kj}^2$  are the component scores at time  $k$ ;  $\sigma_i \in \mathbb{R}$  is estimated variance of the score variable. For different fault types, the data of pneumatic valve in normal working state are collected to establish the principal component model, and different principal component models are established for each fault type. When the pneumatic valve begins to degenerate, the health index  $T^2$  and SPE gradually deviate from the principal component model, and when the deviation is too large, it indicates the failure of the pneumatic valve. The threshold equation of health index  $T^2$  can be represented as:

$$T_\alpha^2 = \frac{A(n-1)}{n-A} F_{A,n-A,\alpha} \quad (4)$$

where  $n$  is the number of the samples,  $A$  is the number of the principal components.  $\alpha$  represents significance level, which value is set at 0.01.  $A$  and  $n-A$  are the confidence of the F distribution.

The threshold equation of health index SPE can be represented as:

$$Q_\alpha = \theta_1 \left( \frac{C_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{\frac{1}{h_0}} \quad (5)$$

where,  $C_\alpha$  is the value of the normal distribution at  $\alpha$  the significance level,  $\theta_i$  and  $h_0$  can be calculated by Eqs. (6) and (7), which represents the eigenvalue of covariance matrix.

$$\theta_i = \sum_{j=A+1}^m \lambda_j^i (i = 1, 2, 3) \quad (6)$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^3} \quad (7)$$

Through the study in this section, the health indicators and threshold values corresponding to the degradation symptom types of each sample can be obtained, which the health indicator will be used to evaluate the current degradation state of the pneumatic valve and make a prediction, and the threshold value of the indicator will be used to judge whether the predicted health indicator has a failure.

According to the  $T^2$  and SPE, the error conditions can be divided into four cases:

1.  $T^2$  and SPE within their threshold limits;
2.  $T^2$  within the threshold limits, but SPE is large than the threshold limits;
3. SPE within the threshold limits, but  $T^2$  is large than the threshold limits;
4.  $T^2$  and SPE are both large than the threshold limits;

Generally, (2), (3) and (4) are failure conditions, while (1) is a normal condition.

### 3.2 Error Prediction Based on Support Vector Regression

In this section, according to the health indicators of different fault types of pneumatic valves obtained, support vector regression models are built respectively for numerical prediction of health indicators. The health indexes of pneumatic valves were constructed as training samples

$$\begin{aligned} D &= \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \\ f(x) &= \omega^T x + b \end{aligned} \quad (8)$$

where  $x \in \mathbb{R}^n$ ,  $y_i \in \mathbb{R}$ .  $f(x)$  is the estimated output of the model.  $\omega$  and  $b$  are respectively the weight and deviation corresponding to the input. But when in the case the distance between  $f(x)$  and  $y$  is greater than  $\varepsilon$ , this case can be considered to be an error. This means that  $\omega$  is as flat as possible.

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m \ell_\varepsilon(f(x_i) - y_i) \quad (9)$$

where  $C$  is the regularization constant, and  $\ell_\varepsilon$  is  $\varepsilon$ -insensitive loss function, which can be formulated as:

$$\ell_\varepsilon(z) = \begin{cases} 0 & , if |z| \leq \varepsilon \\ |z| - \varepsilon, & if |z| > \varepsilon \end{cases} \quad (10)$$

where  $z = y_i - \omega x + b$ . According to Eq. (9), the  $\xi_i$  and  $\hat{\xi}_i$  are used as slack variable. Based on Eqs. (10), (11) can be formulated as:

$$\min_{\omega, b, \xi_i, \hat{\xi}_i} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m (\xi_i + \hat{\xi}_i) \quad (11)$$

where  $-\varepsilon - \xi_i \leq z \leq \varepsilon + \hat{\xi}_i$ ,  $\xi_i \geq 0$ ,  $\hat{\xi}_i \geq 0$  ( $i = 1, 2, \dots, m$ ). According to the Lagrange multiplier and optimal constraints, the kernel method of support vector regression model can be expressed as Eq. (12):

$$f(x) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) \kappa(x, x_i) + b \quad (12)$$

where  $\kappa(x, x_i) = \phi(x_i)^T \phi(x)$  is the kernel function. The commonly used kernel functions of support vector regression algorithms include linear kernel function polynomial kernel function radial basis kernel function and Sigmoid kernel function. The radial basis kernel function has a good performance when dealing with regression problems. The radial basis kernel function is selected as the kernel function of the support vector regression prediction model.

As the degradation process of pneumatic valves generally takes a long time, it is very difficult to obtain a large number of complete life cycle data. Therefore, in order to fully apply the training samples and improve the accuracy of the model, the idea of time sliding window is used to construct the prediction target artificially. The degradation process data of a pneumatic valve is slide-wise intercepted into multiple sample sets in the form of window, and each sample set includes training data and prediction indexes. In order to better train the model, different data sample sets are intercepted from the training set according to time, and each sample set includes training number samples and test samples. The support vector regression algorithm is used for multiple training, which can effectively improve the accuracy of the model.

## 4 Simulation Results and Discussions

In this section, the simulation indicators are introduced, then the joint simulation model fault acquisition method is described, and finally, it is revealed the time for fault occurrence by predicting health indicators based on degradation data.

### 4.1 Failure Prediction Evaluation Index

The prediction of failure is to predict future health indicators based on existing health indicators. In this paper, Mean Squared Error (MSE) and R Squared ( $R^2$ ) are selected as the evaluation indexes of the pneumatic valve failure prediction model, which are calculated as shown in Eqs. (13) and (14).

$$MSE = \frac{1}{m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2 \quad (13)$$

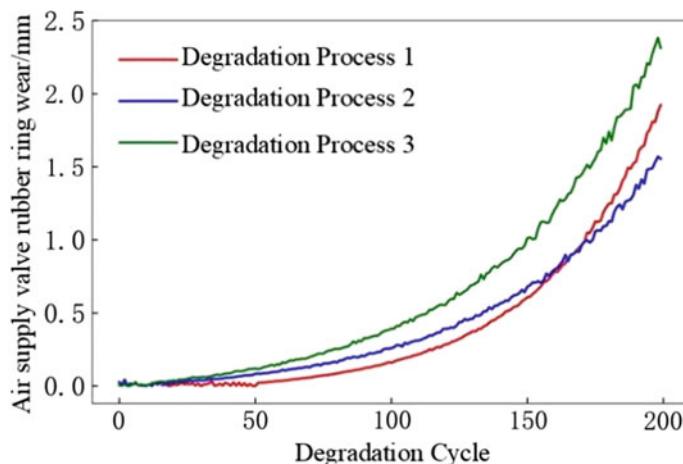
$$\begin{aligned}
 R^2 &= 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\bar{y} - y^{(i)})^2} = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2 / m}{\sum_i (\bar{y} - y^{(i)})^2 / m} \\
 &= 1 - \frac{MSE(\hat{y}, y)}{Var(y)}
 \end{aligned} \tag{14}$$

According to Eqs. (13) and (14), the size of regression evaluation index MSE is related to the dimension of the model. MSE increases with larger dimensions. With consistent dimensions, the decrease in MSE will result in better model performance. The regression evaluation index  $R^2$  removes the dimensionality of the model and returns an accuracy between 0 and 1. Training models with a higher  $R^2$  will have better accuracy.

## 4.2 Preparation for Fault Prediction Data

In order to obtain the process data with fault occurrence, a joint simulation model of AMESim and Matlab was built, and Simulink was used to dynamically adjust the real-time parameters of a part of the relay valve to obtain the fault process data by artificially giving the degradation process curve of the part. When the real-time parameters are given, the model parameters are the same for each cycle, and the model parameters are different between cycles.

In Fig. 2, three degradation curves represent three different degradation processes. We model the different degradation processes for the rubber ring wear in relay



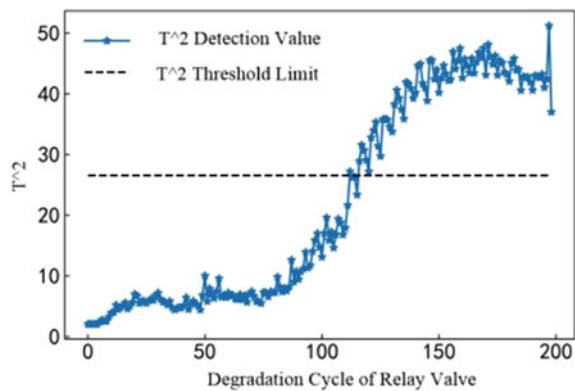
**Fig. 2** The degradation process relay valve supply air valve rubber ring wear failure parameters change curve

valve supply air valves and demonstrate how the proposed failure prediction method accurately predicts the time of failure when the form of degradation is uncertain.

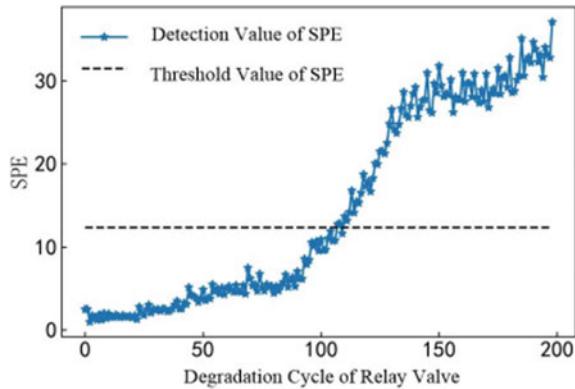
### 4.3 Extracting Health Indicators by Principal Component Analysis

The principal component models are established for the degradation process of rubber ring wear fault of relay valve air supply valve to extract health indicators. Since the establishment of the principal component model needs to test whether the original spatial data obey the normal distribution, K-S test is used to prove that the original spatial data obey the normal distribution, and the principal component analysis method is used to extract the health indicators  $T^2$  and SPE from the original space.

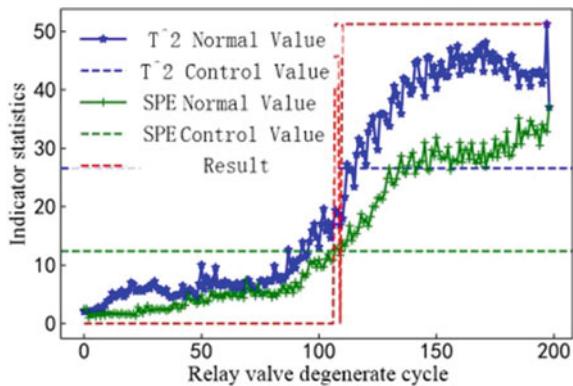
**Fig. 3** Health index  $T^2$  of relay valve



**Fig. 4** Health index SPE of relay valve



**Fig. 5** Fault judgment results



In Figs. 3 and 4, the health indicators  $T^2$  and SPE are represented by the blue curve, and represents the control limits of the two indicators are represented by the black dotted line. The fault judgment results are shown in Fig. 4.

In Fig. 5, the index  $T^2$  is represented by the blue curve, the control limit of  $T^2$  is represented by the blue dotted line, the index SPE is represented by the green curve, the control limit of SPE is represented by the green dotted line, and the fault detection results are represented by the red dotted line. When the red dotted line is equal to 0, it means that the relay valve has no fault. When the red dotted line is not equal to 0, it means that the relay valve has fault. It can be seen from the figure that the relay valve fails after the 105th cycle.

#### 4.4 Result Analysis

The grid search method is used to select the super parameter value of support vector regression model. The value range and step size of super parameter are selected according to experience, and then the grid search method is used to search each super parameter combination within the value range according to step size.

It can be seen from Table 1 that the R-squared values of the first two training samples are relatively large, and the prediction results are relatively good, while the

**Table 1** SVR Evaluation of fault prediction results

Evaluating indicator	R-Squared		MSE	
Health indicators	$T^2$	SPE	$T^2$	SPE
Sample1	0.93357006	0.96067862	16.08351596	6.15154239
Sample2	0.48019985	0.91665023	44.06930541	23.42579304
Sample3	0.74403404	0.90005641	43.07602770	76.77761936
Sample4	0.48877394	0.88889388	20.67797548	64.13940336

R-squared values of  $T^2$  health indicators of the last two training samples are relatively small, indicating that the prediction results are relatively poor. On the whole, the method proposed in this paper has good performance in both fault prediction results and fault prediction accuracy, and can effectively realize the fault prediction of pneumatic valve.

## 5 Conclusion

This paper proposes a novel data driven fault prognosis method for pneumatic valves in train braking system. Two health indicators, i.e.,  $T^2$  and SPE, are extracted through PCA method, which are further used to train a fault prognosis model based on SVR. The proposed method is validated on a semi-physical simulation verification platform of DK-2 braking system, results show that the proposed fault prognosis model can estimate the expected time of pneumatic valve fault time accurately.

## References

1. Soo-Ho, J., Boseong, S., Hyunseok, O., Byeng, D.Y., Dongki, L.: Model-based fault detection method for coil burnout in solenoid valves subjected to dynamic thermal loading. *IEEE Access* **8**, 70387–70400 (2020)
2. Jianyong, Z., Jingxian, D., Shun, P., Guo, H., Tiefeng, Z., Jialiang, L.: Fault feature extraction of relay valve leakage based on test for standard EMU brake system. *DEStech Trans Eng Technol Res* (2017)
3. Yuan, X., Ying, L., Qunxiong, Z.: Multivariate time delay analysis based local KPCA fault prognosis approach for nonlinear processes. *Chin. J. Chem. Eng.* **24**(10), 1413–1422 (2016)
4. Daigle, M.: Real-time prognostics of a rotary valve actuator. In: Annual Conference of the Prognostics and Health Management Society 2015, San Diego, CA, United States (2015)
5. Chetan, S.K., Matthew, D., Kai, G.: Implementation of prognostic methodologies to cryogenic propellant loading testbed. In: IEEE International Automatic Testing Conference 2013, vol. 2013, pp. 1–7. IEEE, Schaumburg, IL, United States (2013)
6. Qing, L., Steven, Y.L.: Degradation trend prognostics for rolling bearing using improved R/S statistic model and fractional Brownian motion approach. *IEEE Access* **6**, 21103–21114 (2017)
7. Mechri, W., Hai-Canh, V., Phuc, D., Klingelschmidt, T., Peysson, F., Theilliol, D.: A study on health diagnosis and prognosis of an industrial diesel motor: hidden Markov models and particle filter approach. *Adv. Solut. Diagnos. Fault Tolerant Control* **635**, 380–389 (2017)
8. Xiaohang, J., Yi, S., Zijun, Q., Yu, W., Tommy, W.S.C.: Anomaly detection and fault prognosis for bearings. *IEEE Trans. Instrum. Measure.* **65**(9), 2046–2054(2016)
9. Min, H., Jinbing, L., Bing, H., Kai, Z.: Fault subspace decomposition and reconstruction theory based Mnline fault prognosis. *Control. Eng. Pract.* **85**, 121–131 (2019)
10. Lotfi, S., Jaouher, B.A., Eric, B., Mohamed, B.: Wind turbine high-speed shaft bearings health prognosis through a spectral Kurtosis-derived indices and SVR. *Appl. Acoust.* **120**, 1–8 (2017)
11. Jichao, H., Zhengpo, W., Yongtao, Y.: Fault prognosis of battery system based on accurate voltage abnormality prognosis using long short-term memory neural networks. *Appl. Energy* **251** (2019)
12. Hack-Eun, K., Andy, C.C.T., Joseph, M., Byeong-Keun, C.: Bearing fault prognosis based on health state probability estimation. *Expert Syst. Appl.* **39**(5), 5200–5213 (2012)

# A RCS Periodicity Extraction Algorithm for Ballistic Target



Chaowei Li , Bing Xie , and Yu Pei

**Abstract** The flight of a ballistic missile can be divided into the boost phase, mid-course flight and the re-entry phase. In the course of missile defense, the identification and interception of mid-course targets are very important. In the mid-course flight of a missile, the warhead is usually surrounded by a large number of targets, which brings great difficulty to the identification of primary targets for early warning radars. This paper establishes a ballistic target echo model, and then proposes a new estimating method for the radar cross-section (RCS) sequence of non-stationary and quasi-periodicity. First, the frequency components of RCS sequence are extracted by the empirical mode decomposition (EMD) with fast Fourier transform. Then, these frequency components are examined by the quantile plots, and finally the correct precession period is extracted based on the results above. The simulation results show that the proposed method extracts the precession period effectively which has a strong anti-noise ability.

**Keywords** Precession period · Empirical mode decomposition · RCS sequence · Quantile plot

## 1 Introduction

The flight of a ballistic missile can be divided into the mid-boost phase and the re-entry phase. The mid-course flight of a ballistic missile has a long distance and a long time, usually accounting for more than 70% of the total flight. In the mid-course flight, the warhead is usually accompanied by a large amount of debris of the booster wreckage, and some decoy-carrying missiles will release various types of decoys in the mid-course flight, which makes it difficult to identify the warhead. When the warhead is flying, in order to hit the target, it will maintain a stable attitude through high speed spin, and produce precession under the interference of the missile

---

C. Li (✉) · B. Xie · Y. Pei  
State Key Laboratory of Astronautic Dynamics, Xi'an, China

body separation, resulting in the periodic change of radar echo. Therefore, extracting precession period of ballistic targets from the radar echo can provide a new way for target recognition.

At present, the research mainly focuses on the cyclic auto-correlation methods (CAUTOC), the cyclic average magnitude difference function (CAMDF) and the analysis of variance. Yao et al. [1] analyzed the common precession period extraction methods under different radar line of sight angles, and proposed a multiple auto-correlation method for period estimation under strong noise. Zhang et al. [2] introduced the convex hull function into the CAMDF to extract the precession period, which had a poor effect in long sequences. Yan et al. [3] combined the Viterbi algorithm and the sum of time-frequency difference squares to estimate the period, which had a poor noise resistance. Zhang et al. [4] proposed a multi-period estimation method, which decomposed the RCS sequence into several periods and then extracted the correct period separately. This method was only applicable to the sequence with simple period components, and did not apply to the targets in mid-course flight.

In this paper, we first analyze the precession model of warhead as well as the law of radar echo, and then propose a new method for estimating the precession period. First, the possible periodic components of original RCS sequence are obtained through empirical mode decomposition (EMD). Then, the rationality of each period is tested by the quantile plot method. Finally, the correct precession period is obtained. Experimental results show that the proposed method can effectively extract the precession period from the RCS sequence, and it has certain robustness to the noise.

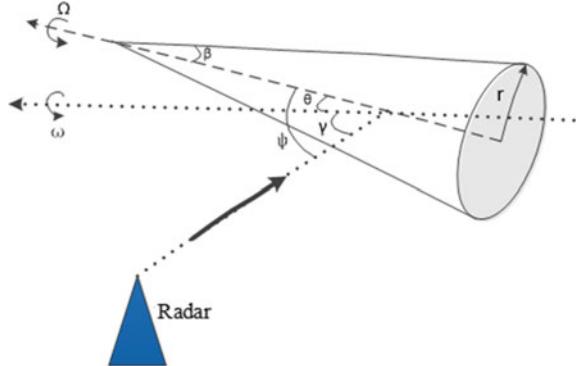
## 2 RCS Analysis of Precession Targets

### 2.1 Target Precession Model

During the mid-flight process, the warhead usually maintains its spin state and precession under the influence of jamming [5]. The precession model of an axisymmetric tapered target is established here (see Fig. 1), in which  $r$  denotes the base radius as well as  $\beta$  denotes the half-cone angle. The direction of precession axis is the same as the direction of warhead velocity. The warhead rotates around its axis at the angular velocity  $\omega$  and precession angle  $\theta (0 < \theta < \frac{\pi}{2})$ , while spins around its axis of symmetry at angular velocity  $\Omega$ . The warhead during mid-flight process is far away from the early warning radar. The target attitude angle  $\psi$  is the angle between radar beam and target symmetry axis, and the average line of sight angle  $\gamma$  is the angle between radar beam and target precession angle. For the axisymmetric type of warhead, its spin has no modulating effect on the radar echo. In this paper, we take the target precession and translation into account.

According to relationship between the angles in Fig. 1, it can be obtained that the attitude angle of the warhead changes with time as follows:

**Fig. 1** Radar detection diagram



$$\psi(t) = \arccos[\cos \theta \cos \gamma(t) - \sin \theta \cos(\omega t + \psi_0) \sin \gamma(t)] \quad (1)$$

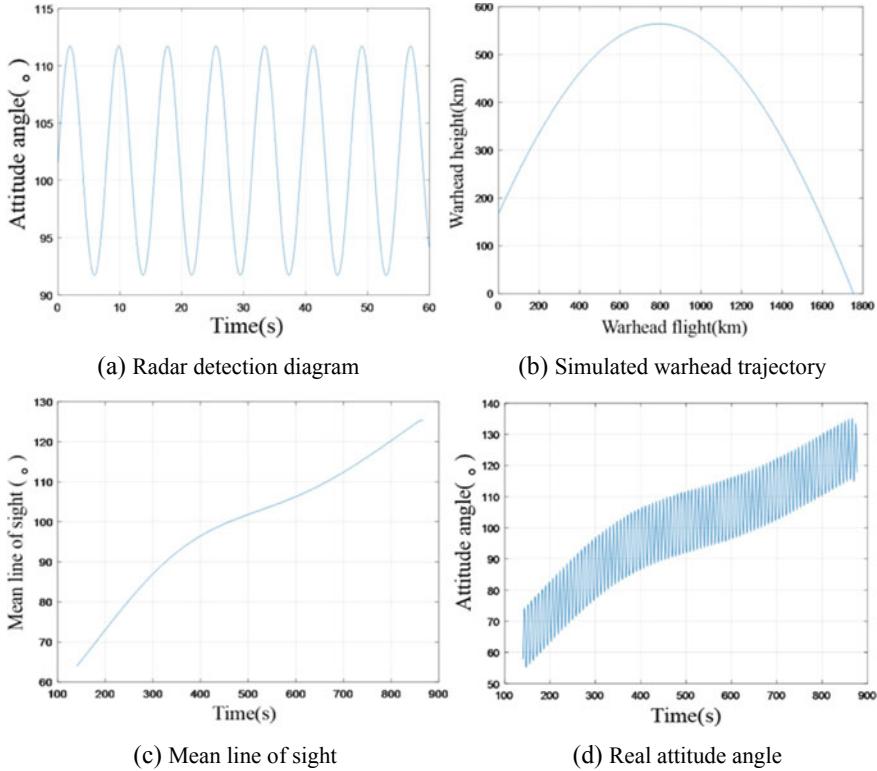
The mean line of sight angle  $\gamma(t)$  is mainly affected by the warhead translation. Here  $\gamma(t)$  is a constant if the warhead does not translate. But in fact, it changes slowly with time during warhead translation. For  $\theta = 10^\circ$ ,  $\omega = 0.8 \text{ rad/s}$ ,  $\gamma(t) = 20^\circ$ , the attitude angle of warhead is shown in Fig. 2a in which changing as cosine periodic curve. A missile launches from  $(33^\circ E, 7^\circ N)$  and lands at the place with the shutdown height of 118 km. Setting the radar deployment located at  $(120^\circ E, 7^\circ N)$ , we simulate the warhead in mid-course flight by STK as shown in Fig. 2b. The vector of the warhead velocity is denoted as  $\vec{v}(t)$  and the vector of radar line-of-sight  $\gamma(t)$  is denoted as  $\vec{r}(t) = \vec{r}_{\text{radar}} - \vec{r}_{\text{target}}$ . While the warhead flies in direction of precession axis, the mean angle of sight is calculated as follows:

$$\gamma(t) = \arccos \frac{\vec{v}(t) \cdot \vec{r}(t)}{|\vec{v}(t)| |\vec{r}(t)|} \quad (2)$$

Substitute the  $\gamma(t)$  above into the Eq. (1), we acquire the attitude angle of warhead during mid-course flight. Based on the trajectory in Fig. 2b, the mean line of sight and real attitude angle are simulated as shown in Fig. 2c and d.

## 2.2 Analysis of Warhead RCS Properties

The radar usually works in high frequency for target recognition. According to the scattering center theory, the radar echo signals in high frequency area can be equalized by sum of several scatter echoes in main parts. The warhead usually contains several scattering centers, and these scattering centers result in the warhead scattering together. The scattering property of axisymmetric tapered targets mainly result from the top and bottom, and the whole RCS can be denoted as [6]:



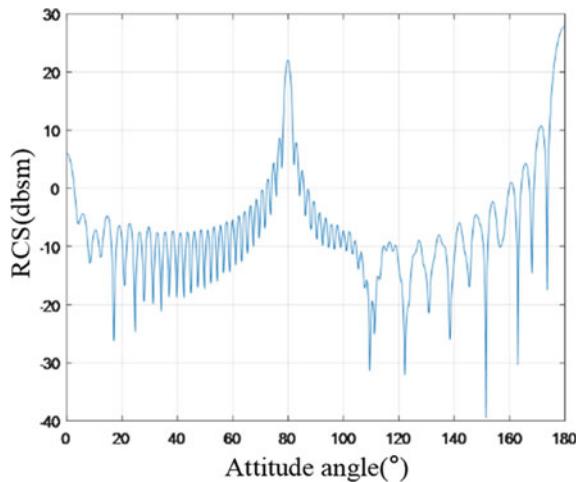
**Fig. 2** The solving process of radar attitude angle

$$\sigma = \left| \sum_{i=1}^3 \sqrt{\sigma_i} e^{j\phi_i} \right|^2 \quad (3)$$

Here  $\sigma_i$  denotes the scattering intensity of the  $i$ th scattering center, which is related to the size, attitude angle and shape of warhead.  $\phi_i$  is the modulation parameter.

The scattering property of warhead in whole attitude angle is shown as Fig. 3 acquired from FEKO, and the parameters of warhead are shown in Table 1.

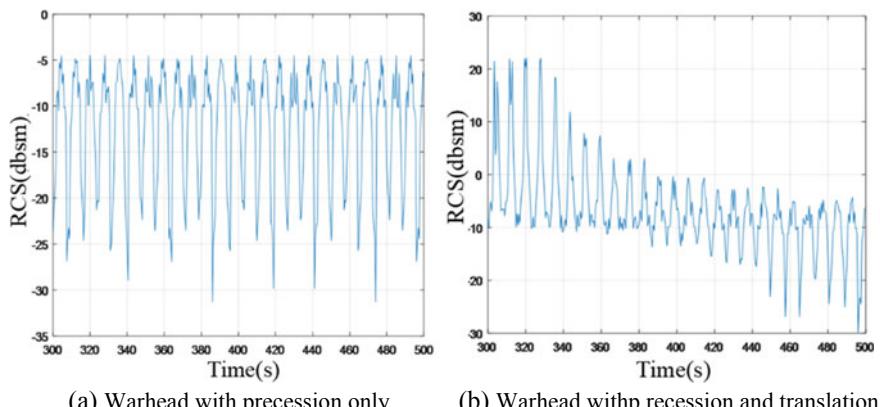
As shown in Fig. 3, the RCS of warhead in whole attitude angle fluctuates obviously, and the same warhead in different attitude angle may have the same echo property. The attitude angle is the connection between static RCS and dynamic RCS, so we can reconstruct dynamic RCS via static RCS. The dynamic RCS resulting from procession is shown as Fig. 4a while the RCS resulting from procession and translation is shown as Fig. 4b. From two figures above, we find RCS of warhead combines non-stationary and non-linearity, which contains several frequencies.



**Fig. 3** Real attitude angle

**Table 1** Parameters of warhead for simulation

Warhead length	Radius of bottom	Radar incidence frequency	Polarization mode
2 m	0.8 m	5.56 GHz	Vertical polarization



**Fig. 4** RCS sequence of the warhead

### 3 Precession Period Estimation

#### 3.1 Empirical Mode Decomposition

The empirical mode decomposition (EMD) method decomposes the signal based on its own property instead of specific basic function [8]. It is more appropriate to a complex nonlinear and non-stationary signal than conventional methods. It will gradually decompose a signal into different frequencies and tendencies, and then we extract several intrinsic mode functions (IMF) with different scale information. The decomposition of RCS sequence  $x(t)$  is as follows:

- (1) First find the local minima and maxima of  $x(t)$ , and then use the local extrema to construct lower and upper envelopes  $S_{\max}$  and  $S_{\min}$ , respectively, of  $x(t)$ ;
- (2) Form the mean of the envelopes  $S_{\text{mean}}$ ;
- (3) Subtract the mean from  $x(t)$  to obtain the residual:  $h_1(t) = x(t) - S_{\text{mean}}(t)$ ;
- (4) Check  $h_1(t)$  and acquire the first IMF if the  $h_1(t)$  satisfies the basic requirement. Otherwise, let  $h_1(t) = x(t)$  and return to step(1 ~ 3) until  $h_i(t)(i = 1,2,3\dots)$  satisfies the basic requirement as follows:

$$SD = \sum_{i=1}^T \frac{|h_{k-1}(t_i) - h_k(t)|^2}{h_{k-1}^2(t_i)} \text{ and } 0.2 < SD < 0.3 \quad (4)$$

Here  $T$  is the length of sequence.

- (5) Let  $r_1(t) = x(t) - h_1(t)$  and regard  $r_1(t)$  as the initial signal, then we repeat above procedures and obtain a series of IMF  $c_i(t)$  ( $i = 1,2,3\dots$ ) and a residue  $r_n(t)$ :

$$x(t) = \sum_{i=1}^n c_i(t) + r_n(t) \quad (5)$$

- (6) If the residue  $r_n(t)$  is a monotonic sequence, stop the decomposition of  $x(t)$ .

Applying the fast Fourier transform algorithm (FFT) to these IMFs above, we obtain the frequency distribution of them. Select the frequency with the highest energy, which is the center frequency, to calculate the period of IMF.

#### 3.2 Extraction of Precession Period

From the chapters above, we can see that warhead RCS has a certain periodicity under the precession modulation, so the RCS sequence in each period should have the same or similar undulation state. Therefore, if we segment the original RCS sequence with correct precession period, each segment should have a similar distribution.

In this paper, we utilize the quantiles-quantiles plot method (Q-Q plot), belonging to non-parametric tests, to determine the final RCS period. Q-Q plot compares the distribution of two samples by quantile plots instead of calculating each value. Non-parametric tests can test the hypothesis of unknown population distribution, such as whether the population distribution is the same or whether it is normal. The points on Q-Q graph will approximate a linear distribution if two samples have the similar distribution, otherwise it will distribute dispersively. An overview of the extraction steps is as follows:

- (1) Segment the original RCS sequence to  $m$  parts according to the those periods, which  $T_i(i = 1,2,3\dots)$  is denoted as  $RCS_{i1}, RCS_{i2} \dots RCS_{im}$ ;
- (2) Draw Q-Q plot  $Q_{ixy}$  for every two segments  $RCS_{ix}, RCS_{iy}$ ;
- (3) The least square fitting method is used for  $Q_{ixy}$  and then calculate the fitting residual  $D_{ixy}$ ;
- (4) Calculate the average fitting residuals  $D_i$  of  $Q_{ixy}$  in period  $T_i$  as follows:

$$D_i = \left( \sum_{x=1}^{m-1} \sum_{y=x+1}^m D_{ixy} \right) / m \quad (6)$$

- (5) Among all the  $T_i(i = 1,2,3\dots)$ , the with the smallest is the correct period.

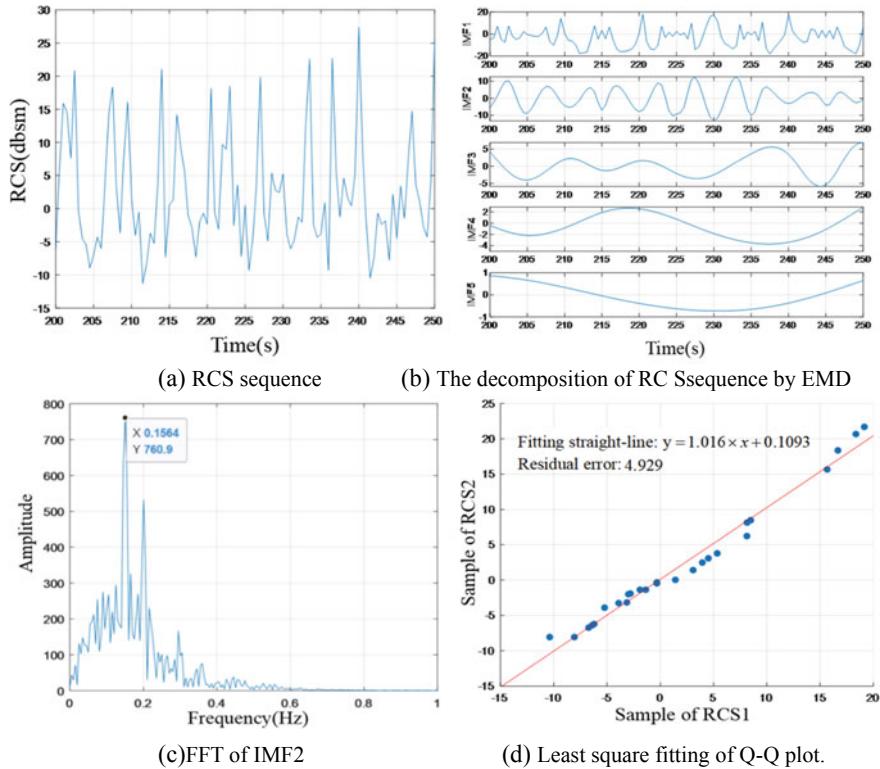
The precession frequency of the warhead is usually a few hertz in reality, which we can base on to select the proper  $T_i(i = 1,2,3\dots)$  in engineering application.

## 4 Experimental Results and Analysis

In this section, the effectiveness of our method is evaluated. The radar transmission frequency is set to 9 GHz and sample frequency is set to 20 Hz. The parameters of warhead are the same as Table 1. Furthermore, the precession angle  $\theta = 10^\circ$  and precession frequency  $f = 0.16$  Hz. The SNR of RCS sequence is set to 10 dB. The original RCS sequence and its decomposition from EMD method are illustrated in Fig. 5a and b. Then adopt FFT to extract the frequency of IMF2 and fit one of the samples using Q-Q plot based on the segments from that frequency as shown in Fig. 5c and d. Check each  $T_i(i = 1,2,3\dots)$  and the results are shown in Table 2.

As shown in Table 2, the fitting residual of IMF2 period is the smallest among all the periods, thus the corresponding frequency 0.156 Hz is regarded as the precession frequency, which is 2.5% lower than the correct frequency. The results from CAMDF [6] and CAUTOC [7] are displayed in Fig. 6a and b for comparison, which are 6 s and 5.5 s respectively. The results are 4.17 and 12.5% higher than the standard, and these methods are prone to false peaks in noise.

The algorithm performances in different noise are shown in Table 3. Due to the original sequence is decomposed into different frequencies by EMD, our algorithm has a stable performance around the correct value with the SNR increased.



**Fig. 5** The solving process of probable frequencies

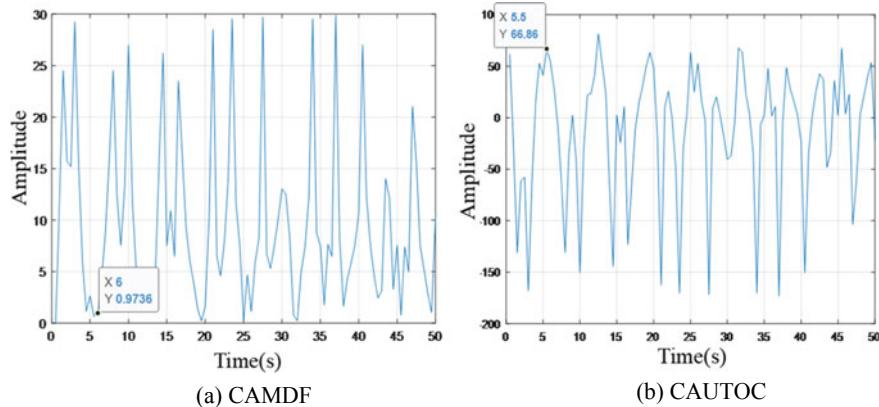
**Table 2** The center frequency of IMFs and the average periods

Modality	The center frequency (Hz)	Average period (s)	The fitting residual of Q-Q plot
IMF1	0.452	2.21	8.34
<b>IMF2</b>	<b>0.156</b>	<b>6.41</b>	<b>4.79</b>
IMF3	0.089	11.24	13.47
IMF4	0.025	40.00	10.23
IMF5	0.018	55.56	7.92

Bold value represents intrinsic mode functions

## 5 Conclusion

In this paper, we first establish the precession model of the warhead, and then analyze the law of RCS of the precession target. Based on the research above, we propose an effective method for precession period extraction. The RCS sequence is decomposed



**Fig. 6** Results from other methods for comparison

**Table 3** The extraction of precession period in different noise

SNR	Correct precession period	Experimental results	Bias (%)
15	4.55	4.42	2.82
18	4.55	4.68	2.92
21	4.55	4.45	2.25
24	4.55	4.47	1.91
27	4.55	4.48	1.57

into different frequency components by EMD and then each precession period is checked by Q-Q plot which finally determine the period. The experimental results demonstrated the effectiveness of the method.

## References

1. Xiaoqiang, Y., Zhiseng, H., Xiangke, G.: The verification of precession period extraction method of ballistic missile based on RCS sequence. *Modern Radar* **39**(11), 63–67 (2017)
  2. Yiyang, Z., Jianjun, G., Gaopeng, L.: Precession period estimation of ballistic target in midcourse. *Telecommun. Eng.* **57**(2), 217–223 (2017)
  3. Zhao, Y., LingLong, G.: Improved micro-motion period estimation method for space targets with compound motion. *J. Electron. Meas. Instrum.* **34**(2), 60–66 (2020)
  4. Ruiguo, Z., Chunyu, L., Yunsheng, H.: An estimation method for multi-period RCS sequences. *Modern Radar* **42**(2), 36–41 (2020)
  5. Wanxing, Z.: Ballistic Missile Radar Target Identification Technology. Publishing House of Electronics Industry, Lin C, Beijing (2011)
  6. Meng, K., Chun-hua, W., Ming, H., Ru-jiang, G.: A study on precession-period extraction method of ballistic targets. *Modern Radar* **32**(11), 29–32 (2010)

7. Huang, N.E., Shen, Z., Long, S.R.: The empirical mode de-composition method and the Hilbert spectrum for non-stationary time series analysis. *Math. Phys. Eng. Sci.* **1971**(454), 903–995 (1998)
8. Weimin, J., Minli, Y., Jianshe, S.: Cyclostationarity of signals and applications. *Modern Radar* **27**(9), 35–39 (2005)