

LNCS 13393

De-Shuang Huang · Kang-Hyun Jo ·
Junfeng Jing · Prashan Premaratne ·
Vitoantonio Bevilacqua · Abir Hussain (Eds.)

Intelligent Computing Theories and Application

18th International Conference, ICIC 2022
Xi'an, China, August 7–11, 2022
Proceedings, Part I



 Springer

Founding Editors

Gerhard Goos

Karlsruhe Institute of Technology, Karlsruhe, Germany

Juris Hartmanis

Cornell University, Ithaca, NY, USA

Editorial Board Members

Elisa Bertino

Purdue University, West Lafayette, IN, USA

Wen Gao

Peking University, Beijing, China

Bernhard Steffen 

TU Dortmund University, Dortmund, Germany

Moti Yung 

Columbia University, New York, NY, USA

More information about this series at <https://link.springer.com/bookseries/558>

De-Shuang Huang · Kang-Hyun Jo ·
Junfeng Jing · Prashan Premaratne ·
Vitoantonio Bevilacqua · Abir Hussain (Eds.)

Intelligent Computing Theories and Application

18th International Conference, ICIC 2022
Xi'an, China, August 7–11, 2022
Proceedings, Part I

Editors

De-Shuang Huang
Tongji University
Shanghai, China

Kang-Hyun Jo
University of Ulsan
Ulsan, Korea (Republic of)

Junfeng Jing
Xi'an Polytechnic University
Xi'an, China

Prashan Premaratne
The University of Wollongong
North Wollongong, NSW, Australia

Vitoantonio Bevilacqua
Polytecnic of Bari
Bari, Italy

Abir Hussain
Liverpool John Moores University
Liverpool, UK

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-13869-0

ISBN 978-3-031-13870-6 (eBook)

<https://doi.org/10.1007/978-3-031-13870-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved 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 Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The International Conference on Intelligent Computing (ICIC) was started to provide an annual forum dedicated to the emerging and challenging topics in artificial intelligence, machine learning, pattern recognition, bioinformatics, and computational biology. It aims to bring together researchers and practitioners from both academia and industry to share ideas, problems, and solutions related to the multifaceted aspects of intelligent computing.

ICIC 2022, held in Xi'an, China, during August 7–11, 2022, constituted the 18th International Conference on Intelligent Computing. It built upon the success of the previous ICIC events held at various locations in China (2005–2008, 2010–2016, 2018–2019, 2021) and in Ulsan, South Korea (2009), Liverpool, UK (2017), and Bari, Italy (2020).

This year, the conference concentrated mainly on the theories, methodologies, and emerging applications of intelligent computing. Its aim was to unify the picture of contemporary intelligent computing techniques as an integral concept that highlights the trends in advanced computational intelligence and bridges theoretical research with applications. Therefore, the theme for this conference was “Advanced Intelligent Computing Technology and Applications”. Papers focused on this theme were solicited, addressing theories, methodologies, and applications in science and technology.

ICIC 2022 received 449 submissions from authors in 21 countries and regions. All papers went through a rigorous peer-review procedure and each paper received at least three review reports. Based on the review reports, the Program Committee finally selected 209 high-quality papers for presentation at ICIC 2022, which are included in three volumes of proceedings published by Springer: two volumes of Lecture Notes in Computer Science (LNCS) and one volume of Lecture Notes in Artificial Intelligence (LNAI).

Among the 449 submissions to the conference were 57 submissions for the six special sessions and nine workshops featured the ICIC this year. All these submissions were reviewed by members from the main Program Committee and 22 high-quality papers were selected for presentation at ICIC 2022 and included in the proceedings based on the topic.

This volume of Lecture Notes in Computer Science (LNCS) includes 66 papers.

The organizers of ICIC 2022, including the EIT Institute for Advanced Study, Xi'an Polytechnic University, Shenzhen University, and the Guangxi Academy of Sciences, made an enormous effort to ensure the success of the conference. We hereby would like to thank the members of the Program Committee and the referees for their collective effort in reviewing and soliciting the papers. In particular, we would like to thank all the authors for contributing their papers. Without the high-quality submissions from the authors, the success of the conference would not have been possible. Finally, we are

especially grateful to the International Neural Network Society and the National Science Foundation of China for their sponsorship.

June 2022

De-Shuang Huang
Kang-Hyun Jo
Junfeng Jing
Prashan Premaratne
Vitoantonio Bevilacqua
Abir Hussain

Organization

General Co-chairs

De-Shuang Huang

Tongji University, China

Haiyan Wang

Xi'an Polytechnic University, China

Program Committee Co-chairs

Kang-Hyun Jo

University of Ulsan, South Korea

Junfeng Jing

Xi'an Polytechnic University, China

Prashan Premaratne

University of Wollongong, Australia

Vitoantonio Bevilacqua

Polytechnic University of Bari, Italy

Abir Hussain

Liverpool John Moores University, UK

Organizing Committee Co-chairs

Pengfei Li

Xi'an Polytechnic University, China

Kaibing Zhang

Xi'an Polytechnic University, China

Lei Zhang

Xi'an Polytechnic University, China

Organizing Committee

Hongwei Zhang

Xi'an Polytechnic University, China

Minqi Li

Xi'an Polytechnic University, China

Zhaoliang Meng

Xi'an Polytechnic University, China

Peng Song

Xi'an Polytechnic University, China

Award Committee Co-chairs

Kyungsook Han

Inha University, South Korea

Valeriya Gribova

Far Eastern Branch of the Russian Academy of Sciences, Russia

Tutorial Co-chairs

Ling Wang

Tsinghua University, China

M. Michael Gromiha

Indian Institute of Technology Madras, India

Publication Co-chairs

Michał Choras	Bydgoszcz University of Science and Technology, Poland
Hong-Hee Lee	University of Ulsan, South Korea
Laurent Heutte	Université de Rouen Normandie, France

Special Session Co-chairs

Yu-Dong Zhang	University of Leicester, UK
Vitoantonio Bevilacqua	Polytechnic University of Bari, Italy
Hee-Jun Kang	University of Ulsan, South Korea

Special Issue Co-chairs

Yoshinori Kuno	Saitama University, Japan
Phalguni Gupta	Indian Institute of Technology Kanpur, India

International Liaison Co-chair

Prashan Premaratne	University of Wollongong, Australia
--------------------	-------------------------------------

Workshop Co-chairs

Jair Cervantes Canales	Autonomous University of Mexico State, Mexico
Chenxi Huang	Xiamen University, China
Dhiya Al-Jumeily	Liverpool John Moores University, UK

Publicity Co-chairs

Chun-Hou Zheng	Anhui University, China
Dhiya Al-Jumeily	Liverpool John Moores University, UK
Jair Cervantes Canales	Autonomous University of Mexico State, Mexico

Sponsors and Exhibits Chair

Qinghu Zhang	Tongji University, China
--------------	--------------------------

Program Committee

Abir Hussain	Liverpool John Moores University, UK
Angelo Ciaramella	Parthenope University of Naples, Italy
Antonino Staiano	Parthenope University of Naples, Italy
Antonio Brunetti	Polytechnic University of Bari, Italy

Bai Xue	Institute of Software, CAS, China
Baitong Chen	Xuzhou No. 1 Peoples Hospital, China
Ben Niu	Shenzhen University, China
Bin Liu	Beijing Institute of Technology, China
Bin Qian	Kunming University of Science and Technology, China
Bin Wang	Anhui University of Technology, China
Bin Yang	Zaozhuang University, China
Bingqiang Liu	Shandong University, China
Binhua Tang	Hohai University, China
Bo Li	Wuhan University of Science and Technology, China
Bo Liu	Academy of Mathematics and Systems Science, CAS, China
Bohua Zhan	Institute of Software, CAS, China
Changqing Shen	Soochow University, China
Chao Song	Harbin Medical University, China
Chenxi Huang	Xiamen University, China
Chin-Chih Chang	Chung Hua University, Taiwan, China
Chunhou Zheng	Anhui University, China
Chunmei Liu	Howard University, USA
Chunquan Li	Harbin Medical University, China
Dah-Jing Jwo	National Taiwan Ocean University, Taiwan, China
Dakshina Ranjan Kisku	National Institute of Technology Durgapur, India
Daowen Qiu	Sun Yat-sen University, China
Dhiya Al-Jumeily	Liverpool John Moores University, UK
Domenico Buongiorno	Politecnico di Bari, Italy
Dong Wang	University of Jinan, China
Dong-Joong Kang	Pusan National University, South Korea
Dunwei Gong	China University of Mining and Technology, China
Eros Gian Pasero	Politecnico di Torino, Italy
Evi Sjukur	Monash University, Australia
Fa Zhang	Institute of Computing Technology, CAS, China
Fabio Stroppa	Stanford University, USA
Fei Han	Jiangsu University, China
Fei Guo	Central South University, China
Fei Luo	Wuhan University, China
Fengfeng Zhou	Jilin University, China
Gai-Ge Wang	Ocean University of China, China
Giovanni Dimauro	University of Bari, Italy
Guojun Dai	Hangzhou Dianzi University, China

Haibin Liu	Beijing University of Technology, China
Han Zhang	Nankai University, China
Hao Lin	University of Electronic Science and Technology of China, China
Haodi Feng	Shandong University, China
Ho-Jin Choi	Korea Advanced Institute of Science and Technology, South Korea
Hong-Hee Lee	University of Ulsan, South Korea
Hongjie Wu	Suzhou University of Science and Technology, China
Hongmin Cai	South China University of Technology, China
Jair Cervantes	Autonomous University of Mexico State, Mexico
Jian Huang	University of Electronic Science and Technology of China, China
Jian Wang	China University of Petroleum (East China), China
Jiangning Song	Monash University, Australia
Jiawei Luo	Hunan University, China
Jieren Cheng	Hainan University, China
Jing Hu	Wuhan University of Science and Technology, China
Jing-Yan Wang	Abu Dhabi Department of Community Development, UAE
Jinwen Ma	Peking University, China
Jin-Xing Liu	Qufu Normal University, China
Ji-Xiang Du	Huaqiao University, China
Joaquin Torres-Sospedra	Universidade do Minho, Portugal
Juan Liu	Wuhan University, China
Junfeng Man	Hunan First Normal University, China
Junfeng Xia	Anhui University, China
Jungang Lou	Huzhou University, China
Junqi Zhang	Tongji University, China
Ka-Chun Wong	City University of Hong Kong, Hong Kong, China
Kanghyun Jo	University of Ulsan, South Korea
Kyungsook Han	Inha University, South Korea
Lejun Gong	Nanjing University of Posts and Telecommunications, China
Laurent Heutte	Université de Rouen Normandie, France
Le Zhang	Sichuan University, China
Lin Wang	University of Jinan, China
Ling Wang	Tsinghua University, China
Li-Wei Ko	National Yang Ming Chiao Tung University, Taiwan, China

Marzio Pennisi	University of Eastern Piedmont, Italy
Michael Gromiha	Indian Institute of Technology Madras, India
Michal Choras	Bydgoszcz University of Science and Technology, Poland
Mine Sarac	Stanford University, USA, and Kadir Has University, Turkey
Mohd Helmy Abd Wahab	Universiti Tun Hussein Onn Malaysia, Malaysia
Na Zhang	Xuzhou Medical University, China
Nicholas Caporusso	Northern Kentucky University, USA
Nicola Altini	Polytechnic University of Bari, Italy
Peng Chen	Anhui University, China
Pengjiang Qian	Jiangnan University, China
Phalguni Gupta	GLA University, India
Ping Guo	Beijing Normal University, China
Prashan Premaratne	University of Wollongong, Australia
Pu-Feng Du	Tianjin University, China
Qi Zhao	University of Science and Technology Liaoning, China
Qingfeng Chen	Guangxi University, China
Qinghua Jiang	Harbin Institute of Technology, China
Quan Zou	University of Electronic Science and Technology of China, China
Rui Wang	National University of Defense Technology, China
Ruiping Wang	Institute of Computing Technology, CAS, China
Saiful Islam	Aligarh Muslim University, India
Seeja K. R.	Indira Gandhi Delhi Technical University for Women, India
Shanfeng Zhu	Fudan University, China
Shanwen Wang	Xijing University, China
Shen Yin	Harbin Institute of Technology, China
Shihua Zhang	Academy of Mathematics and Systems Science, CAS, China
Shihua Zhang	Wuhan University of Science and Technology, China
Shikui Tu	Shanghai Jiao Tong University, China
Shitong Wang	Jiangnan University, China
Shixiong Zhang	Xidian University, China
Shunren Xia	Zhejiang University, China
Sungshin Kim	Pusan National University, South Korea
Surya Prakash	Indian Institute Technology Indore, India
Takashi Kuremoto	Nippon Institute of Technology, Japan
Tao Zeng	Guangzhou Laboratory, China

Tatsuya Akutsu	Kyoto University, Japan
Tieshan Li	University of Electronic Science and Technology of China, China
Valeriya Gribova	Institute of Automation and Control Processes, Far Eastern Branch of the Russian Academy of Sciences, Russia
Vincenzo Randazzo	Politecnico di Torino, Italy
Waqas Haider Bangyal	University of Gujrat, Pakistan
Wei Chen	Chengdu University of Traditional Chinese Medicine, China
Wei Jiang	Nanjing University of Aeronautics and Astronautics, China
Wei Peng	Kunming University of Science and Technology, China
Wei Wei	Tencent Technology, Norway
Wei-Chiang Hong	Asia Eastern University of Science and Technology, Taiwan, China
Weidong Chen	Shanghai Jiao Tong University, China
Weihong Deng	Beijing University of Posts and Telecommunications, China
Weixiang Liu	Shenzhen University, China
Wen Zhang	Huazhong Agricultural University, China
Wenbin Liu	Guangzhou University, China
Wen-Sheng Chen	Shenzhen University, China
Wenzheng Bao	Xuzhou University of Technology, China
Xiangtao Li	Jilin University, China
Xiaodi Li	Shandong Normal University, China
Xiaofeng Wang	Hefei University, China
Xiao-Hua Yu	California Polytechnic State University, USA
Xiaoke Ma	Xidian University, China
Xiaolei Zhu	Anhui Agricultural University, China
Xiaoli Lin	Wuhan University of Science and Technology, China
Xiaoqi Zheng	Shanghai Normal University, China
Xin Yin	Laxco Inc., USA
Xin Zhang	Jiangnan University, China
Xinguo Lu	Hunan University, China
Xingwen Liu	Southwest Minzu University, China
Xiujuan Lei	Shaanxi Normal University, China
Xiwei Liu	Tongji University, China
Xiyuan Chen	Southeast University, China
Xuequn Shang	Northwestern Polytechnical University, China

Xuesong Wang	China University of Mining and Technology, China
Xuesong Yan	China University of Geosciences, China
Xu-Qing Tang	Jiangnan University, China
Yan-Rui Ding	Jiangnan University, China
Yansen Su	Anhui University, China
Yi Gu	Jiangnan University, China
Yi Xiong	Shanghai Jiao Tong University, China
Yizhang Jiang	Jiangnan University, China
Yong-Quan Zhou	Guangxi University for Nationalities, China
Yonggang Lu	Lanzhou University, China
Yoshinori Kuno	Saitama University, Japan
Yu Xue	Huazhong University of Science and Technology, China
Yuan-Nong Ye	Guizhou Medical University, China
Yu-Dong Zhang	University of Leicester, UK
Yue Ming	Beijing University of Posts and Telecommunications, China
Yunhai Wang	Shandong University, China
Yupei Zhang	Northwestern Polytechnical University, China
Yushan Qiu	Shenzhen University, China
Zhanheng Chen	Shenzhen University, China
Zhan-Li Sun	Anhui University, China
Zhen Lei	Institute of Automation, CAS, China
Zhendong Liu	Shandong Jianzhu University, China
Zhenran Jiang	East China Normal University, China
Zhenyu Xuan	University of Texas at Dallas, USA
Zhi-Hong Guan	Huazhong University of Science and Technology, China
Zhi-Ping Liu	Shandong University, China
Zhiqiang Geng	Beijing University of Chemical Technology, China
Zhongqiu Zhao	Hefei University of Technology, China
Zhu-Hong You	Northwestern Polytechnical University, China
Zhuo Wang	Hangzhou Dianzi University, China
Zuguo Yu	Xiangtan University, China

Contents – Part I

Evolutionary Computing and Learning

Evolutionary Game Analysis of Suppliers Considering Quality Supervision of the Main Manufacturer	3
<i>Tiaojuan Han, Jianfeng Lu, and Hao Zhang</i>	
Multi-party Evolution Stability Analysis of Electric Vehicles- Microgrid Interaction Mechanism	13
<i>Haitong Guo, Hao Zhang, Jianfeng Lu, Rong Zeng, and Tiaojuan Han</i>	
An Efficient Multi-objective Evolutionary Algorithm for a Practical Dynamic Pickup and Delivery Problem	27
<i>Junchuang Cai, Qingling Zhu, Qiuzhen Lin, Jianqiang Li, Jianyong Chen, and Zhong Ming</i>	
An Efficient Evaluation Mechanism for Evolutionary Reinforcement Learning	41
<i>Xiaoqiang Wu, Qingling Zhu, Qiuzhen Lin, Jianqiang Li, Jianyong Chen, and Zhong Ming</i>	
A Mixed-Factor Evolutionary Algorithm for Multi-objective Knapsack Problem	51
<i>Yanlian Du, Zejing Feng, and Yijun Shen</i>	
NSLS with the Clustering-Based Entropy Selection for Many-Objective Optimization Problems	68
<i>Zhaobin Ma, Bowen Ding, and Xin Zhang</i>	
Tunicate Swarm Algorithm Based Difference Variation Flower Pollination Algorithm	80
<i>Chuchu Yu, Huajuan Huang, and Xiuxi Wei</i>	
A Multi-strategy Improved Fireworks Optimization Algorithm	97
<i>Pengcheng Zou, Huajuan Huang, and Xiuxi Wei</i>	
A New Fitness-Landscape-Driven Particle Swarm Optimization	112
<i>Xuying Ji, Feng Zou, Debao Chen, and Yan Zhang</i>	
Neighborhood Combination Strategies for Solving the Bi-objective Max-Bisection Problem	123
<i>Rong-Qiang Zeng and Matthieu Basseur</i>	

Neural Networks

Rolling Bearing Fault Diagnosis Based on Model Migration	135
<i>Yuchen Xing and Hui Li</i>	
Yak Management Platform Based on Neural Network and Path Tracking	147
<i>Yunfan Hu</i>	
Stability Analysis of Hopfield Neural Networks with Conformable Fractional Derivative: M-matrix Method	159
<i>Chang-bo Yang, Sun-yan Hong, Ya-qin Li, Hui-meи Wang, and Yan Zhu</i>	
Artificial Neural Networks for COVID-19 Forecasting in Mexico: An Empirical Study	168
<i>C. M. Castorena, R. Alejo, E. Rendón, E. E. Granda-Gutiérrez, R. M. Valdovinos, and G. Miranda-Piña</i>	
Multi-view Robustness-Enhanced Weakly Supervised Semantic Segmentation	180
<i>Yu Sang, Shi Li, and Yanfei Peng</i>	
Rolling Bearing Fault Diagnosis Based on Graph Convolution Neural Network	195
<i>Yin Zhang and Hui Li</i>	
Research on Bearing Fault Feature Extraction Based on Graph Wavelet	208
<i>Xin Li and Hui Li</i>	
Correntrogram: A Robust Method for Optimal Frequency Band Selection to Bearing Fault Detection	221
<i>Hui Li, Ruijuan Wang, and Yonghui Xie</i>	
Semidefinite Relaxation Algorithm for Source Localization Using Multiple Groups of TDOA Measurements with Distance Constraints	233
<i>Tao Zhang, Wuyi Yang, and Yu Zhang</i>	

Pattern Recognition

Quasi Fourier Descriptor for Affine Invariant Features	245
<i>Chengyun Yang, Lei Lu, Lei Zhang, Yu Tian, and Zhang Chen</i>	
A New PM2.5 Concentration Predication Study Based on CNN-LSTM Parallel Integration	258
<i>Chaoxue Wang, Zhenbang Wang, Fan Zhang, and Yuhang Pan</i>	

Deep Discriminant Non-negative Matrix Factorization Method for Image Clustering	267
<i>Kexin Xie, Wen-Sheng Chen, and Binbin Pan</i>	
A Feature Extraction Algorithm for Enhancing Graphical Local Adaptive Threshold	277
<i>Shaoshao Wang, Aihua Zhang, and Han Wang</i>	
Person Re-identification Based on Transform Algorithm	292
<i>Lei Xie, Chao Wang, Xiaoyong Yu, Aihua Zheng, and Guolong Chen</i>	
Modified Lightweight U-Net with Attention Mechanism for Weld Defect Detection	306
<i>Lei Huang, Shanwen Zhang, Liang Li, Xiulin Han, Ruijiang Li, Hongbo Zhang, and Shaoqing Sun</i>	
Handwritten Chemical Equations Recognition Based on Lightweight Networks	317
<i>Xiao-Feng Wang, Zhi-Huang He, Zhi-Ze Wu, Yun-Sheng Wei, Kai Wang, and Le Zou</i>	
Illumination Invariant Face Recognition Using Directional Gradient Maps	330
<i>Guang Yi Chen, Wenfang Xie, and Adam Krzyzak</i>	
News Video Description Based on Template Generation and Entity Insertion ...	339
<i>Qiyang Yuan, Pengjun Zhai, Dulei Zheng, and Yu Fang</i>	
Local Feature for Visible-Thermal PReID Based on Transformer	352
<i>Quanyi Pu, Changan Yuan, Hongjie Wu, and Xingming Zhao</i>	
A Hardware Implementation Method of Radar Video Scanning Transformation Based on Dual FPGA	363
<i>Naizhao Yu, Xiao Min, and Liang Zhao</i>	
Image Processing	
An Image Binarization Segmentation Method Combining Global and Local Threshold for Uneven Illumination Image	379
<i>Jin-Wu Wang, Daiwei Xie, and Zhenmin Dai</i>	
Optimization of Vessel Segmentation Using Genetic Algorithms	391
<i>Jared Cervantes, Dalia Luna, Jair Cervantes, and Farid García-Lamont</i>	
Graph-Based Anomaly Detection via Attention Mechanism	401
<i>Yangming Yu, Zhiyong Zha, Bo Jin, Geng Wu, and Chenxi Dong</i>	

A Classification Algorithm Based on Discriminative Transfer Feature Learning for Early Diagnosis of Alzheimer's Disease	412
<i>Xinchun Cui, Yonglin Liu, Jianzong Du, Qinghua Sheng, Xiangwei Zheng, Yue Feng, Liying Zhuang, Xiuming Cui, Jing Wang, and Xiaoli Liu</i>	
A Systematic Study for the Effects of PCA on Hyperspectral Imagery Denoising	420
<i>Guang Yi Chen and Wen Fang Xie</i>	
Two-Channel VAE-GAN Based Image-To-Video Translation	430
<i>Shengli Wang, Mulin Xieshi, Zhangpeng Zhou, Xiang Zhang, Xujie Liu, Zeyi Tang, Yuxing Dai, Xuexin Xu, and Pingyuan Lin</i>	
High-Voltage Tower Nut Detection and Positioning System Based on Binocular Vision	444
<i>Zhiyu Cheng, YiHua Luo, JinFeng Zhang, Zhiwen Gong, Lei Sun, and Lang Xu</i>	
Palmprint Recognition Using the Combined Method of BEMD and WCB-NNSC	456
<i>Li Shang, Yuze Zhang, and Zhan-li Sun</i>	
Palmprint Feature Extraction Utilizing WTA-ICA in Contourlet Domain	464
<i>Li Shang, Yuze Zhang, and Zhan-li Sun</i>	
Blockwise Feature-Based Registration of Deformable Medical Images	472
<i>Su Wai Tun, Takashi Komuro, and Hajime Nagahara</i>	
Measuring Shape and Reflectance of Real Objects Using a Handheld Camera	483
<i>Shwe Yee Win, Zar Zar Tun, Seiji Tsunezaki, and Takashi Komuro</i>	
Image-to-Video Translation Using a VAE-GAN with Refinement Network	494
<i>Shengli Wang, Mulin Xieshi, Zhangpeng Zhou, Xiang Zhang, Xujie Liu, Zeyi Tang, Jianbing Xiahou, Pingyuan Lin, Xuexin Xu, and Yuxing Dai</i>	
Joint Semantic Segmentation and Object Detection Based on Relational Mask R-CNN	506
<i>Yanni Zhang, Hui Xu, Jingxuan Fan, Miao Qi, Tao Liu, and Jianzhong Wang</i>	
Stitching High Resolution Notebook Keyboard Surface Based on Halcon Calibration	522
<i>Gang Lv, Hao Zhao, Zuchang Ma, Yining Sun, and Fudong Nian</i>	

An Improved NAMLab Image Segmentation Algorithm Based on the Earth Moving Distance and the CIEDE2000 Color Difference Formula	535
<i>Yunping Zheng, Yuan Xu, Shengjie Qiu, Wenqiang Li, Guichuang Zhong, Mingyi Chen, and Mudar Sarem</i>	
An Improved NAMLab Algorithm Based on CIECDE2000 Color Difference Formula and Gabor Filter for Image Segmentation	549
<i>Yunping Zheng, Shengjie Qiu, Jiehao Huang, Yuan Xu, Zirui Zou, and Pengcheng Sun</i>	
An Improved Block Truncation Coding Using Rectangular Non-symmetry and Anti-packing Model	564
<i>Yunping Zheng, Yuan Xu, Jinjun Kuang, and Mudar Sarem</i>	
Image Super-Resolution Reconstruction Based on MCA and ICA Denoising	579
<i>Weiguo Yang, Bin Yang, Jing Li, and Zhongyu Sun</i>	
Image Representation Based on Overlapping Rectangular NAM and Binary Bit-Plane Decomposition	589
<i>Yunping Zheng, Yuan Xu, Jinjun Kuang, and Mudar Sarem</i>	
Information Security	
Research on the Rule of Law in Network Information Security Governance	605
<i>Pei Zhaobin and Yu Yixiao</i>	
Legal Analysis of the Right to Privacy Protection in the Age of Artificial Intelligence	617
<i>Sun Xin, Pei Zhaobin, and Qu Jing</i>	
An Intrusion Detection Method Fused Deep Learning and Fuzzy Neural Network for Smart Home	627
<i>Xiangdong Hu, Qin Zhang, Xi Yang, and Liu Yang</i>	
A High Performance Intrusion Detection System Using LightGBM Based on Oversampling and Undersampling	638
<i>Hao Zhang, Lina Ge, and Zhe Wang</i>	
Research on the Current Situation and Improvement Countermeasures of Farmers' Information Security Literacy Based on New Media	653
<i>Haiyu Wang</i>	

A Torque-Current Prediction Model Based on GRU for Circumferential Rotation Force Feedback Device	663
<i>Zekang Qiu, Jianhui Zhao, Chudong Shan, Wenyuan Zhao, Tingbao Zhang, and Zhiyong Yuan</i>	
Development and Application of Augmented Reality System for Part Assembly Based on Assembly Semantics	673
<i>Yingxin Wang, Jianfeng Lu, Zeyuan Lin, Lai Dai, Junxiong Chen, and Luyao Xia</i>	
Research on Augmented Reality Assisted Material Delivery System in Digital Workshop	685
<i>Zhaojia Li, Hao Zhang, Jianfeng Lu, and Luyao Xia</i>	
Biomedical Informatics Theory and Methods	
Safety and Efficacy of Short-Term vs. Standard Periods Dual Antiplatelet Therapy After New-Generation Drug-Eluting Stent Implantation: A Meta-analysis	701
<i>Xiaohua Gao, Xiaodan Bi, Jimpeng Yang, and Meili Cheng</i>	
Automated Diagnosis of Vertebral Fractures Using Radiographs and Machine Learning	726
<i>Li-Wei Cheng, Hsin-Hung Chou, Kuo-Yuan Huang, Chin-Chiang Hsieh, Po-Lun Chu, and Sun-Yuan Hsieh</i>	
Cost and Care Insight: An Interactive and Scalable Hierarchical Learning System for Identifying Cost Saving Opportunities	739
<i>Yuan Zhang, David Koepke, Bibo Hao, Jing Mei, Xu Min, Rachna Gupta, Rajashree Joshi, Fiona McNaughton, Zhan-Heng Chen, Bo-Wei Zhao, Lun Hu, and Pengwei Hu</i>	
A Sub-network Aggregation Neural Network for Non-invasive Blood Pressure Prediction	753
<i>Xinghui Zhang, Chunhou Zheng, Peng Chen, Jun Zhang, and Bing Wang</i>	
Integrating Knowledge Graph and Bi-LSTM for Drug-Drug Interaction Predication	763
<i>Shanwen Zhang, Changqing Yu, and Cong Xu</i>	
A 3D Medical Image Segmentation Framework Fusing Convolution and Transformer Features	772
<i>Fazhan Zhu, Jiaxing Lv, Kun Lu, Wenyan Wang, Hongshou Cong, Jun Zhang, Peng Chen, Yuan Zhao, and Ziheng Wu</i>	

COVID-19 Classification from Chest X-rays Based on Attention and Knowledge Distillation	787
<i>Jiaxing Lv, Fazhan Zhu, Kun Lu, Wenyan Wang, Jun Zhang, Peng Chen, Yuan Zhao, and Ziheng Wu</i>	
Using Deep Learning to Predict Transcription Factor Binding Sites Based on Multiple-omics Data	799
<i>Youhong Xu, Changan Yuan, Hongjie Wu, and Xingming Zhao</i>	
Non-invasive Haemoglobin Prediction Using Nail Color Features: An Approach of Dimensionality Reduction	811
<i>Sunanda Das, Abhishek Kesarwani, Dakshina Ranjan Kisku, and Mamata Dalui</i>	
Author Index	825

Contents – Part II

Biomedical Data Modeling and Mining

A Comparison Study of Predicting lncRNA-Protein Interactions via Representative Network Embedding Methods	3
<i>Guoqing Zhao, Pengpai Li, and Zhi-Ping Liu</i>	
GATSDCD: Prediction of circRNA-Disease Associations Based on Singular Value Decomposition and Graph Attention Network	14
<i>Mengting Niu, Abd El-Latif Hesham, and Quan Zou</i>	
Anti-breast Cancer Drug Design and ADMET Prediction of ERA Antagonists Based on QSAR Study	28
<i>Wentao Gao, Ziyi Huang, Hao Zhang, and Jianfeng Lu</i>	
Real-Time Optimal Scheduling of Large-Scale Electric Vehicles Based on Non-cooperative Game	41
<i>Rong Zeng, Hao Zhang, Jianfeng Lu, Tiaojuan Han, and Haitong Guo</i>	
TBC-Unet: U-net with Three-Branch Convolution for Gliomas MRI Segmentation	53
<i>Yongpu Yang, Haitao Gan, and Zhi Yang</i>	
Drug–Target Interaction Prediction Based on Graph Neural Network and Recommendation System	66
<i>Peng Lei, Changan Yuan, Hongjie Wu, and Xingming Zhao</i>	
NSAP: A Neighborhood Subgraph Aggregation Method for Drug–Disease Association Prediction	79
<i>Qiqi Jiao, Yu Jiang, Yang Zhang, Yadong Wang, and Junyi Li</i>	
Comprehensive Evaluation of BERT Model for DNA-Language for Prediction of DNA Sequence Binding Specificities in Fine-Tuning Phase ...	92
<i>Xianbao Tan, Changan Yuan, Hongjie Wu, and Xingming Zhao</i>	
Identification and Evaluation of Key Biomarkers of Acute Myocardial Infarction by Machine Learning	103
<i>Zhenrun Zhan, Tingting Zhao, Xiaodan Bi, Jinpeng Yang, and Pengyong Han</i>	

Glioblastoma Subtyping by ImmunoGenomics	116
<i>Yanran Li, Chandrasekhar Gopalakrishnan, Jian Wang, Rajasekaran Ramalingam, Caixia Xu, and Pengyong Han</i>	
Functional Analysis of Molecular Subtypes with Deep Similarity Learning Model Based on Multi-omics Data	126
<i>Shuhui Liu, Zhang Yupei, and Xuequn Shang</i>	
Predicting Drug-Disease Associations by Self-topological Generalized Matrix Factorization with Neighborhood Constraints	138
<i>Xiaoguang Li, Qiang Zhang, Zonglan Zuo, Rui Yan, Chunhou Zheng, and Fa Zhang</i>	
Intelligent Computing in Computational Biology	
iEnhancer-BERT: A Novel Transfer Learning Architecture Based on DNA-Language Model for Identifying Enhancers and Their Strength	153
<i>Hanyu Luo, Cheng Chen, Wenyu Shan, Pingjian Ding, and Lingyun Luo</i>	
GCNMFCDA: A Method Based on Graph Convolutional Network and Matrix Factorization for Predicting circRNA-Disease Associations	166
<i>Dian-Xiao Wang, Cun-Mei Ji, Yu-Tian Wang, Lei Li, Jian-Cheng Ni, and Bin Li</i>	
Prediction of MiRNA-Disease Association Based on Higher-Order Graph Convolutional Networks	181
<i>Zhengtao Zhang, Pengyong Han, Zhengwei Li, Ru Nie, and Qiankun Wang</i>	
SCDF: A Novel Single-Cell Classification Method Based on Dimension-Reduced Data Fusion	196
<i>Chujie Fang and Yuanyuan Li</i>	
Research on the Potential Mechanism of Rhizoma Drynariae in the Treatment of Periodontitis Based on Network Pharmacology	207
<i>Caixia Xu, Xiaokun Yang, Zhipeng Wang, Pengyong Han, Xiaoguang Li, and Zhengwei Li</i>	
Predicting Drug-Disease Associations via Meta-path Representation Learning based on Heterogeneous Information Net works	220
<i>Meng-Long Zhang, Bo-Wei Zhao, Lun Hu, Zhu-Hong You, and Zhan-Heng Chen</i>	
An Enhanced Graph Neural Network Based on the Tissue-Like P System	233
<i>Dongyi Li and Xiyu Liu</i>	

Cell Classification Based on Stacked Autoencoder for Single-Cell RNA Sequencing	245
<i>Rong Qi, Chun-Hou Zheng, Cun-Mei Ji, Ning Yu, Jian-Cheng Ni, and Yu-Tian Wang</i>	
A Novel Cuprotosis-Related Gene Signature Predicts Survival Outcomes in Patients with Clear-Cell Renal Cell Carcinoma	260
<i>Zhenrun Zhan, Pengyong Han, Xiaodan Bi, Jinpeng Yang, and Tingting Zhao</i>	
Identification of miRNA-lncRNA Underlying Interactions Through Representation for Multiplex Heterogeneous Network	270
<i>Jiren Zhou, Zhuhong You, Xuequn Shang, Rui Niu, and Yue Yun</i>	
ACNN: Drug-Drug Interaction Prediction Through CNN and Attention Mechanism	278
<i>Weizhi Wang and Hongbo Liu</i>	
Elucidating Quantum Semi-empirical Based QSAR, for Predicting Tannins' Anti-oxidant Activity with the Help of Artificial Neural Network	289
<i>Chandrasekhar Gopalakrishnan, Caixia Xu, Yanran Li, Vinutha Anandhan, Sanjay Gangadharan, Meshach Paul, Chandra Sekar Ponnusamy, Rajasekaran Ramalingam, Pengyong Han, and Zhengwei Li</i>	
Drug-Target Interaction Prediction Based on Transformer	302
<i>Junkai Liu, Tengsheng Jiang, Yaoyao Lu, and Hongjie Wu</i>	
Protein-Ligand Binding Affinity Prediction Based on Deep Learning	310
<i>Yaoyao Lu, Junkai Liu, Tengsheng Jiang, Shixuan Guan, and Hongjie Wu</i>	
Computational Genomics and Biomarker Discovery	
Position-Defined CpG Islands Provide Complete Co-methylation Indexing for Human Genes	319
<i>Ming Xiao, Ruiying Yin, Pengbo Gao, Jun Yu, Fubo Ma, Zichun Dai, and Le Zhang</i>	
Predicting the Subcellular Localization of Multi-site Protein Based on Fusion Feature and Multi-label Deep Forest Model	334
<i>Hongri Yang, Qingfang Meng, Yuehui Chen, and Lianxin Zhong</i>	

Construction of Gene Network Based on Inter-tumor Heterogeneity for Tumor Type Identification	345
<i>Zhensheng Sun, Junliang Shang, Hongyu Duan, Jin-Xing Liu, Xikui Liu, Yan Li, and Feng Li</i>	
A Novel Synthetic Lethality Prediction Method Based on Bidirectional Attention Learning	356
<i>Fengxu Sun, Xinguo Lu, Guanyuan Chen, Xiang Zhang, Kaibao Jiang, and Jinxin Li</i>	
A Novel Trajectory Inference Method on Single-Cell Gene Expression Data ...	364
<i>Daoxu Tang, Xinguo Lu, Kaibao Jiang, Fengxu Sun, and Jinxin Li</i>	
Bioinformatic Analysis of Clear Cell Renal Carcinoma via ATAC-Seq and RNA-Seq	374
<i>Feng Chang, Zhenqiong Chen, Caixia Xu, Hailei Liu, and Pengyong Han</i>	
The Prognosis Model of Clear Cell Renal Cell Carcinoma Based on Allograft Rejection Markers	383
<i>Hailei Liu, Zhenqiong Chen, Chandrasekhar Gopalakrishnan, Rajasekaran Ramalingam, Pengyong Han, and Zhengwei li</i>	
Membrane Protein Amphiphilic Helix Structure Prediction Based on Graph Convolution Network	394
<i>Baoli Jia, Qingfang Meng, Qiang Zhang, and Yuehui Chen</i>	
The CNV Predict Model in Esophagus Cancer	405
<i>Yun Tian, Caixia Xu, Lin Li, Pengyong Han, and Zhengwei Li</i>	
TB-LNPs: A Web Server for Access to Lung Nodule Prediction Models	415
<i>Huaichao Luo, Ning Lin, Lin Wu, Ziru Huang, Ruiling Zu, and Jian Huang</i>	
Intelligent Computing in Drug Design	
A Targeted Drug Design Method Based on GRU and TopP Sampling Strategies	423
<i>Jinglu Tao, Xiaolong Zhang, and Xiaoli Lin</i>	
KGAT: Predicting Drug-Target Interaction Based on Knowledge Graph Attention Network	438
<i>Zhenghao Wu, Xiaolong Zhang, and Xiaoli Lin</i>	

MRLDTI: A Meta-path-Based Representation Learning Model for Drug-Target Interaction Prediction	451
<i>Bo-Wei Zhao, Lun Hu, Peng-Wei Hu, Zhu-Hong You, Xiao-Rui Su, Dong-Xu Li, Zhan-Heng Chen, and Ping Zhang</i>	
Single Image Dehazing Based on Generative Adversarial Networks	460
<i>Mengyun Wu and Bo Li</i>	
K-Nearest Neighbor Based Local Distribution Alignment	470
<i>Yang Tian and Bo Li</i>	
A Video Anomaly Detection Method Based on Sequence Recognition	481
<i>Lei Yang and Xiaolong Zhang</i>	
Drug-Target Binding Affinity Prediction Based on Graph Neural Networks and Word2vec	496
<i>Minghao Xia, Jing Hu, Xiaolong Zhang, and Xiaoli Lin</i>	
Drug-Target Interaction Prediction Based on Attentive FP and Word2vec	507
<i>Yi Lei, Jing Hu, Ziyu Zhao, and Siyi Ye</i>	
Unsupervised Prediction Method for Drug-Target Interactions Based on Structural Similarity	517
<i>Xinyuan Zhang, Xiaoli Lin, Jing Hu, and Wenquan Ding</i>	
Drug-Target Affinity Prediction Based on Multi-channel Graph Convolution	533
<i>Hang Zhang, Jing Hu, and Xiaolong Zhang</i>	
An Optimization Method for Drug-Target Interaction Prediction Based on RandSAS Strategy	547
<i>Huimin Xiang, AoXing Li, and Xiaoli Lin</i>	
A Novel Cuprotosis-Related lncRNA Signature Predicts Survival Outcomes in Patients with Glioblastoma	556
<i>Hongyu Sun, Xiaohui Li, Jin Yang, Yi Lyu, Pengyong Han, and Jinping Zheng</i>	
Arbitrary Voice Conversion via Adversarial Learning and Cycle Consistency Loss	569
<i>Jie Lian, Pingyuan Lin, Yuxing Dai, and Guilin Li</i>	
MGVC: A Mask Voice Conversion Using Generating Adversarial Training	579
<i>Pingyuan Lin, Jie Lian, and Yuxing Dai</i>	

Covid-19 Detection by Wavelet Entropy and Genetic Algorithm	588
<i>Jia-Ji Wan, Shu-Wen Chen, Rayan S. Cloutier, and Hui-Sheng Zhu</i>	
COVID-19 Diagnosis by Wavelet Entropy and Particle Swarm Optimization ...	600
<i>Jia-Ji Wang</i>	
Theoretical Computational Intelligence and Applications	
An Integrated GAN-Based Approach to Imbalanced Disk Failure Data	615
<i>Shuangshuang Yuan, Peng Wu, Yuehui Chen, Liqiang Zhang, and Jian Wang</i>	
Disk Failure Prediction Based on Transfer Learning	628
<i>Guangfu Gao, Peng Wu, Hui Li, and Tianze Zhang</i>	
Imbalanced Disk Failure Data Processing Method Based on CTGAN	638
<i>Jingbo Jia, Peng Wu, Kai Zhang, and Ji Zhong</i>	
SID ² T: A Self-attention Model for Spinal Injury Differential Diagnosis	650
<i>Guan Wang, Yulin Wu, Qinghua Sun, Bin Yang, and Zhaona Zheng</i>	
Predicting Protein-DNA Binding Sites by Fine-Tuning BERT	663
<i>Yue Zhang, Yuehui Chen, Baitong Chen, Yi Cao, Jiazi Chen, and Hanhan Cong</i>	
i6mA-word2vec: A Newly Model Which Used Distributed Features for Predicting DNA N6-Methyladenine Sites in Genomes	670
<i>Wenzhen Fu, Yixin Zhong, Baitong Chen, Yi Cao, Jiazi Chen, and Hanhan Cong</i>	
Oxides Classification with Random Forests	680
<i>Kai Xiao, Baitong Chen, Wenzheng Bao, and Honglin Cheng</i>	
Protein Sequence Classification with LetNet-5 and VGG16	687
<i>Zheng Tao, Zhen Yang, Baitong Chen, Wenzheng Bao, and Honglin Cheng</i>	
SeqVec-GAT: A Golgi Classification Model Based on Multi-headed Graph Attention Network	697
<i>Jianan Sui, Yuehui Chen, Baitong Chen, Yi Cao, Jiazi Chen, and Hanhan Cong</i>	
Classification of S-succinylation Sites of Cysteine by Neural Network	705
<i>Tong Meng, Yuehui Chen, Baitong Chen, Yi Cao, Jiazi Chen, and Hanhan Cong</i>	

E. coli Proteins Classification with Naive Bayesian	715
<i>Yujun Liu, Jiaxin Hu, Yue Zhou, Wenzheng Bao, and Honglin Cheng</i>	
COVID-19 and SARS Virus Function Sites Classification with Machine Learning Methods	722
<i>Hongdong Wang, Zizhou Feng, Baitong Chen, Wenhao Shao, Zijun Shao, Yumeng Zhu, and Zhuo Wang</i>	
Identification of Protein Methylation Sites Based on Convolutional Neural Network	731
<i>Wenzheng Bao, Zuo Wang, and Jian Chu</i>	
Image Repair Based on Least Two-Way Generation Against the Network	739
<i>Juxi Hu and Honglin Cheng</i>	
Prediction of Element Distribution in Cement by CNN	747
<i>Xin Zhao, Yihan Zhou, Jianfeng Yuan, Bo Yang, Xu Wu, Dong Wang, Pengwei Guan, and Na Zhang</i>	
An Ensemble Framework Integrating Whole Slide Pathological Images and miRNA Data to Predict Radiosensitivity of Breast Cancer Patients	757
<i>Chao Dong, Jie Liu, Wenhui Yan, Mengmeng Han, Lijun Wu, Junfeng Xia, and Yannan Bin</i>	
Bio-ATT-CNN: A Novel Method for Identification of Glioblastoma	767
<i>Jinling Lai, Zhen Shen, and Lin Yuan</i>	
STE-COVIDNet: A Multi-channel Model with Attention Mechanism for Time Series Prediction of COVID-19 Infection	777
<i>Hongjian He, Xinwei Lu, Dingkai Huang, and Jiang Xie</i>	
KDPCnet: A Keypoint-Based CNN for the Classification of Carotid Plaque	793
<i>Bindong Liu, Wu Zhang, and Jiang Xie</i>	
Multi-source Data-Based Deep Tensor Factorization for Predicting Disease-Associated miRNA Combinations	807
<i>Sheng You, Zihan Lai, and Jiawei Luso</i>	
Author Index	823

Contents – Part III

Fuzzy Theory and Algorithms

An Incremental Approach Based on Hierarchical Classification in Multikernel Fuzzy Rough Sets Under the Variation of Object Set	3
<i>Wei Fan, Chunlin He, Anping Zeng, and Ke Lin</i>	
A Clustering Method Based on Improved Density Estimation and Shared Nearest Neighbors	18
<i>Ying Guan, Yaru Li, Bin Li, and Yonggang Lu</i>	
Bagging-AdaTSK: An Ensemble Fuzzy Classifier for High-Dimensional Data	32
<i>Guangdong Xue, Bingjie Zhang, Xiaoling Gong, and Jian Wang</i>	
Some Results on the Dominance Relation Between Conjunctions and Disjunctions	44
<i>Lizhu Zhang and Gang Li</i>	
Robust Virtual Sensors Design for Linear Systems	55
<i>Alexey Zhirabok, Alexander Zuev, Vladimir Filaretov, Changan Yuan, Alexander Protcenko, and Kim Chung Il</i>	
Clustering Analysis in the Student Academic Activities on COVID-19 Pandemic in Mexico	67
<i>G. Miranda-Piña, R. Alejo, E. Rendón, E. E. Granda-Gutiérrez, R. M. Valdovinos, and F. del Razo-López</i>	
Application of Stewart Platform as a Haptic Device for Teleoperation of a Mobile Robot	80
<i>Duc-Vinh Le and CheolKeun Ha</i>	
Geometric Parameters Calibration Method for Multilink Manipulators	93
<i>Anton Gubankov, Dmitry Yukhimets, Vladimir Filaretov, and Changan Yuan</i>	
A Kind of PWM DC Motor Speed Regulation System Based on STM32 with Fuzzy-PID Dual Closed-Loop Control	106
<i>Wang Lu, Zhang Zaitian, Cheng Xuwei, Ren Haoyu, Chen Jianzhou, Qiu Fengqi, Yan Zitong, Zhang Xin, and Zhang Li</i>	

Machine Learning and Data Mining

Research on Exchange and Management Platform of Enterprise Power Data Unification Summit	117
<i>Yangming Yu, Zhiyong Zha, Bo Jin, Geng Wu, and Chenxi Dong</i>	
Application of Deep Learning Autoencoders as Features Extractor of Diabetic Foot Ulcer Images	129
<i>Abbas Saad Alatrany, Abir Hussain, Saad S. J. Alatrany, and Dhiya Al-Jumaily</i>	
MPCNN with Knowledge Augmentation: A Model for Chinese Text Classification	141
<i>Xiaozeng Zhang and Ailian Fang</i>	
An Improved Mobilenetv2 for Rust Detection of Angle Steel Tower Bolts Based on Small Sample Transfer Learning	150
<i>Zhiyu Cheng, Jun Liu, and Jinfeng Zhang</i>	
Generate Judge-View of Online Dispute Resolution Based on Pretrained-Model Method	162
<i>Qinhua Huang and Weimin Ouyang</i>	
An Effective Chinese Text Classification Method with Contextualized Weak Supervision for Review Autograting	170
<i>Yupei Zhang, Md Shahedul Islam Khan, Yaya Zhou, Min Xiao, and Xuequn Shang</i>	
Comparison of Subjective and Physiological Stress Levels in Home and Office Work Environments	183
<i>Matthew Harper, Fawaz Ghali, and Wasiq Khan</i>	
Cross Distance Minimization for Solving the Nearest Point Problem Based on Scaled Convex Hull	198
<i>Qiangkui Leng, Erjie Jiao, Yuqing Liu, Jiamei Guo, and Ying Chen</i>	
Nut Recognition and Positioning Based on YOLOv5 and RealSense	209
<i>JinFeng Zhang, TianZhong Zhang, Jun Liu, Zhiwen Gong, and Lei Sun</i>	
Gait Identification Using Hip Joint Movement and Deep Machine Learning	220
<i>Luke Topham, Wasiq Khan, Dhiya Al-Jumeily, Atif Waraich, and Abir Hussain</i>	

Study on Path Planning of Multi-storey Parking Lot Based on Combined Loss Function	234
<i>Zhongtian Hu, Jun Yan, Yuli Wang, Changsong Yang, Qiming Fu, Weizhong Lu, and Hongjie Wu</i>	
A Systematic Review of Distributed Deep Learning Frameworks for Big Data	242
<i>Francesco Berloco, Vitoantonio Bevilacqua, and Simona Colucci</i>	
Efficient Post Event Analysis and Cyber Incident Response in IoT and E-commerce Through Innovative Graphs and Cyberthreat Intelligence Employment	257
<i>Rafał Kozik, Marek Pawlicki, Mateusz Szczepański, Rafał Renk, and Michał Chorąś</i>	
Federated Sparse Gaussian Processes	267
<i>Xiangyang Guo, Daging Wu, and Jinwen Ma</i>	
Classification of Spoken English Accents Using Deep Learning and Speech Analysis	277
<i>Zaid Al-Jumaili, Tarek Bassiouny, Ahmad Alanezi, Wasiq Khan, Dhiya Al-Jumeily, and Abir Jaafar Hussain</i>	
A Stable Community Detection Approach for Large-Scale Complex Networks Based on Improved Label Propagation Algorithm	288
<i>Xiangtao Chen and Meijie Zhao</i>	
An Effective Method for Yemeni License Plate Recognition Based on Deep Neural Networks	304
<i>Hamdan Taleb, Zhipeng Li, Changan Yuan, Hongjie Wu, Xingming Zhao, and Fahd A. Ghanem</i>	
Topic Analysis of Public Welfare Microblogs in the Early Period of the COVID-19 Epidemic Based on LDA Model	315
<i>Ji Li and Yujun Liang</i>	
Intelligent Computing in Computer Vision	
Object Detection Networks and Mixed Reality for Cable Harnesses Identification in Assembly Environment	331
<i>Yixiong Wei, Hongqi Zhang, Hongqiao Zhou, Qianhao Wu, and Zihan Niu</i>	
Improved YOLOv5 Network with Attention and Context for Small Object Detection	341
<i>Tian-Yu Zhang, Jun Li, Jie Chai, Zhong-Qiu Zhao, and Wei-Dong Tian</i>	

Inverse Sparse Object Tracking via Adaptive Representation	353
<i>Jian-Xun Mi, Yun Gao, and Renjie Li</i>	
A Sub-captions Semantic-Guided Network for Image Captioning	367
<i>Wei-Dong Tian, Jun-jun Zhu, Shuang Wu, Zhong-Qiu Zhao, Yu-Zheng Zhang, and Tian-yu Zhang</i>	
A Novel Gaze Detection Method Based on Local Feature Fusion	380
<i>Juan Li, Yahui Dong, Hui Xu, Hui Sun, and Miao Qi</i>	
Vehicle Detection, Classification and Counting on Highways - Accuracy Enhancements	394
<i>Prashan Premaratne, Rhys Blackridge, and Mark Lee</i>	
Image Dehazing Based on Deep Multiscale Fusion Network and Continuous Memory Mechanism	409
<i>Qiang Li, Zhihua Xie, Sha Zong, and Guodong Liu</i>	
Improved YOLOv5s Model for Vehicle Detection and Recognition	423
<i>Xingmin Lu and Wei Song</i>	
Garbage Classification Detection Model Based on YOLOv4 with Lightweight Neural Network Feature Fusion	435
<i>Xiao-Feng Wang, Jian-Tao Wang, Li-Xiang Xu, Ming Tan, Jing Yang, and Yuan-yan Tang</i>	
Detection of Personal Protective Equipment in Factories: A Survey and Benchmark Dataset	448
<i>Zhiyang Liu, Thomas Weise, and Zhize Wu</i>	
Intelligent Control and Automation	
A Novel IoMT System for Pathological Diagnosis Based on Intelligent Mobile Scanner and Whole Slide Image Stitching Method	463
<i>Peng Jiang, Juan Liu, Di Xiao, Baochuan Pang, Zongjie Hao, and Dehua Cao</i>	
Deep Reinforcement Learning Algorithm for Permutation Flow Shop Scheduling Problem	473
<i>Yuanyuan Yang, Bin Qian, Rong Hu, and Dacheng Zhang</i>	
Model Predictive Control for Voltage Regulation in Bidirectional Boost Converter	484
<i>Duy-Long Nguyen, Huu-Cong Vu, Quoc-Hoan Tran, and Hong-Hee Lee</i>	

Fed-MT-ISAC: Federated Multi-task Inverse Soft Actor-Critic for Human-Like NPCs in the Metaverse Games	492
<i>Fangze Lin, Wei Ning, and Zhengrong Zou</i>	
Development of AUV Two-Loop Sliding Control System with Considering of Thruster Dynamic	504
<i>Filaretov Vladimir, Yukhimets Dmitry, and Changan Yuan</i>	
An Advanced Terminal Sliding Mode Controller for Robot Manipulators in Position Tracking Problem	518
<i>Anh Tuan Vo, Thanh Nguyen Truong, Hee-Jun Kang, and Tien Dung Le</i>	
An Observer-Based Fixed Time Sliding Mode Controller for a Class of Second-Order Nonlinear Systems and Its Application to Robot Manipulators	529
<i>Thanh Nguyen Truong, Anh Tuan Vo, Hee-Jun Kang, and Tien Dung Le</i>	
A Robust Position Tracking Strategy for Robot Manipulators Using Adaptive Second Order Sliding Mode Algorithm and Nonsingular Sliding Mode Control	544
<i>Tan Van Nguyen, Cheolkeun Ha, Huy Q. Tran, Dinh Hai Lam, and Nguyen Thi Hoa Cuc</i>	
Intelligent Data Analysis and Prediction	
A Hybrid Daily Carbon Emission Prediction Model Combining CEEMD, WD and LSTM	557
<i>Xing Zhang and Wensong Zhang</i>	
A Hybrid Carbon Price Prediction Model Based on VMD and ELM Optimized by WOA	572
<i>Xing Zhang and Wensong Zhang</i>	
A Comparative Study of Autoregressive and Neural Network Models: Forecasting the GARCH Process	589
<i>Firuz Kamalov, Ikhlaas Gurrib, Sherif Moussa, and Amril Nazir</i>	
A Novel DCT-Based Video Steganography Algorithm for HEVC	604
<i>Si Liu, Yunxia Liu, Cong Feng, and Hongguo Zhao</i>	
Dynamic Recurrent Embedding for Temporal Interaction Networks	615
<i>Qilin Liu, Xiaobo Zhu, Changgan Yuan, Hongje Wu, and Xinming Zhao</i>	

Deep Spatio-Temporal Attention Network for Click-Through Rate Prediction	626
<i>Xin-Lu Li, Peng Gao, Yuan-Yuan Lei, Le-Xuan Zhang, and Liang-Kuan Fang</i>	
A Unified Graph Attention Network Based Framework for Inferring circRNA-Disease Associations	639
<i>Cun-Mei Ji, Zhi-Hao Liu, Li-Juan Qiao, Yu-Tian Wang, and Chun-Hou Zheng</i>	
Research on the Application of Blockchain Technology in the Evaluation of the “Five Simultaneous Development” Education System	654
<i>Xian-hong Xu, Feng-yang Sun, and Yu-qing Zheng</i>	
Blockchain Adoption in University Archives Data Management	662
<i>Cong Feng and Si Liu</i>	
A Novel Two-Dimensional Histogram Shifting Video Steganography Algorithm for Video Protection in H.265/HEVC	672
<i>Hongguo Zhao, Yunxia Liu, and Yonghao Wang</i>	
Problems and Countermeasures in the Construction of Intelligent Government Under the Background of Big Data	684
<i>ZhaoBin Pei and Ying Wang</i>	
Application of Auto-encoder and Attention Mechanism in Raman Spectroscopy	698
<i>Yunyi Bai, Mang Xu, and Pengjiang Qian</i>	
Remaining Useful Life Prediction Based on Improved LSTM Hybrid Attention Neural Network	709
<i>Mang Xu, Yunyi Bai, and Pengjiang Qian</i>	
Medical Image Registration Method Based on Simulated CT	719
<i>Xuqing Wang, Yanan Su, Ruoyu Liu, Qianhui Qu, Hao Liu, and Yi Gu</i>	
Research on Quantitative Optimization Method Based on Incremental Optimization	729
<i>Ying Chen, Youjun Huang, and Lichao Gao</i>	
An Improved Waste Detection and Classification Model Based on YOLOV5 ...	741
<i>Fan Hu, Pengjiang Qian, Yizhang Jiang, and Jian Yao</i>	

An Image Compression Method Based on Compressive Sensing and Convolution Neural Network for Massive Imaging Flow Cytometry Data	755
<i>Long Cheng and Yi Gu</i>	
Intelligent Computing and Optimization	
Optimization Improvement and Clustering Application Based on Moth-Flame Algorithm	769
<i>Lvyang Ye, Huajuan Huang, and Xiuxi Wei</i>	
Application of Improved Fruit Fly Optimization Algorithm in Three Bar Truss	785
<i>Dao Tao, Xiuxi Wei, and Huajuan Huang</i>	
Adaptive Clustering by Fast Search and Find of Density Peaks	802
<i>Yuanyuan Chen, Lina Ge, Guifen Zhang, and Yongquan Zhou</i>	
A “Push-Pull” Workshop Logistics Distribution Under Single Piece and Small-Lot Production Mode	814
<i>Mengxia Xu, Hao Zhang, Xue Wang, and Jianfeng Lu</i>	
Greedy Squirrel Search Algorithm for Large-Scale Traveling Salesman Problems	830
<i>Chenghao Shi, Zhonghua Tang, Yongquan Zhou, and Qifang Luo</i>	
Multiple Populations-Based Whale Optimization Algorithm for Solving Multicarrier NOMA Power Allocation Strategy Problem	846
<i>Zhiwei Liang, Qifang Luo, and Yongquan Zhou</i>	
Complex-Valued Crow Search Algorithm for 0–1 KP Problem	860
<i>Yan Shi, Yongquan Zhou, Qifang Luo, and Huajuan Huang</i>	
Discrete Artificial Electric Field Optimization Algorithm for Graph Coloring Problem	876
<i>Yixuan Yu, Yongquan Zhou, Qifang Luo, and Xiuxi Wei</i>	
Automatic Shape Matching Using Improved Whale Optimization Algorithm with Atomic Potential Function	891
<i>Yuanfei Wei, Ying Ling, Qifang Luo, and Yongquan Zhou</i>	
Author Index	907

Evolutionary Computing and Learning



Evolutionary Game Analysis of Suppliers Considering Quality Supervision of the Main Manufacturer

Tiaojuan Han, Jianfeng Lu^(✉), and Hao Zhang

CIMS Research Center, Tongji University, Shanghai 201804, China
lujianfeng@tongji.edu.cn

Abstract. In the manufacturing process of high-end equipment, the quality of parts is an important factor affecting the quality of the final product, so it is necessary for the main manufacturer to supervise the quality of suppliers' parts. There is little research on quality supervision in “the main manufacturer-multiple suppliers” mode. Considering the quality supervision of the main manufacturer, an evolutionary game model of quality improvement between two suppliers is established based on evolutionary game theory. Then, the stability of the game equilibrium point is discussed based on the stability criterion of Jacobian matrix. Finally, the impacts of penalty coefficient of the main manufacturer on suppliers' strategies are analyzed through numerical simulation. The result illustrates that the increase in penalty coefficient can motivate suppliers to improve the quality of parts, and the result provides theoretical guidance for the main manufacturer to supervise suppliers.

Keywords: Suppliers · Evolutionary game · Quality supervision

1 Introduction

High-end equipment is a kind of product or system with high technology, such as marine engineering equipment. The manufacturing process of high-end equipment involves multiple enterprises and fields. Suppliers process raw materials and product parts. Then, the main manufacturer assembles parts, and delivers final products to users. The manufacturing characteristic is “main manufacturer-multiple suppliers”. The quality of parts is one of important factors affecting the quality of the product. The main manufacturer expects high-quality parts from suppliers to ensure the quality of the final product, but suppliers improving the quality of parts pay additional costs, and are unwilling to improve the quality of parts. Therefore, regulatory mechanism of the main manufacturer is considered to encourage suppliers to improve the quality of parts [1]. Considering the main manufacturer and suppliers with bounded rationality, the study is based on evolutionary game theory.

This paper considers the quality supervision of the main manufacturer, and establishes an evolutionary game between supplier A and supplier B. Finally, the impact of penalty coefficient on the evolution process of suppliers is analyzed through numerical simulation. The results provide supervision suggestions for manufacturers.

2 Related Work

Evolutionary game has been widely used in various fields.

- (1) some scholars studied platform governance through evolutionary game. Wang et al. [2] constructed an evolutionary game model to analyze the platform supervision. Weng et al. [3] constructed an evolutionary game model of the car-hailing platform.
- (2) Evolutionary game is also used to analyze the strategy evolution of supply and demand relationship. Li et al. [4] constructed a tripartite evolutionary game model of local government, reporting company and agency, and analyzed the impacts of variables on strategy selection. Cheng et al. [5] established an asymmetric evolutionary game model of the government, developers and consumers, and analyzed the impact of parameters on the balance of supply and demand. Shan et al. [6] analyzed the tripartite evolutionary game process of manufacturing enterprises, digital service platform and consumers. Yang et al. [7] constructed a tripartite evolutionary game model of the government, enterprises and the public in responsible innovation. Wang et al. [8] constructed a tripartite evolution model of the government, innovative supply enterprise and demand enterprise considering green technology innovation. He et al. [9] constructed an evolutionary game model of enterprise cooperation in supply and demand network.
- (3) Scholars have studied the evolutionary process between manufacturers and suppliers, such as green innovation between manufacturers and suppliers [10, 11]. Li et al. [12] studied the game evolution process of financial risk cooperation between suppliers and manufacturers. Ma et al. [13] analyzed the price game process between manufacturers and suppliers.

In summary, evolutionary game theory has been widely used in platform governance, etc. However, there is little research on strategy evolution of “main manufacturer–multiple suppliers” based on evolutionary game theory. This paper considers the quality improvement of parts and the main manufacturer’ supervision to study the evolutionary game between the main manufacturer and suppliers.

3 Two-Party Evolutionary Game Between Suppliers

3.1 Model Assumptions and Payoff Matrix

As the core enterprise of the manufacturing process, the main manufacturer not only assembles parts, but also coordinates and supervises suppliers. Based on evolutionary game, the paper analyzes the strategy evolution process between parts’ suppliers. When parts’ suppliers do not improve the quality of parts, they are punished by the main manufacturer.

Hypothesis 1. As players, supplier A and supplier B are bounded rationality. there are two strategies: (improve the quality of parts(I), not improve the quality of parts (NI)). If suppliers do not improve the quality of parts, expenditure cost is production cost of parts. To improve the quality of parts, suppliers introduce advanced equipment and technology, and pay additional costs. The probability of supplier A improving the

quality of parts is $X, 0 \leq X \leq 1$. When $X = 1$, supplier A improves the quality of parts. When $X = 0$ supplier A does not improve the quality of parts. The probability of supplier B improving the quality of parts is $Y, 0 \leq Y \leq 1$. When $Y = 1$, supplier B improves the quality of parts. When $Y = 0$, supplier B does not improve the quality of parts.

Hypothesis 2. G_i means suppliers' revenues of providing parts, $G_i > 0, i = A, B$; C_{iH} means suppliers' production cost of improving the quality of parts, $C_{iH} \geq 0$; C_{iL} is suppliers' production cost of not improving the quality of parts, $C_{iL} \geq 0$. a indicates penalty coefficient of the main manufacturer for not improving the quality of parts, $0 \leq a \leq 1$. C_i is the main manufacturer's penalty for the supplier i if one of suppliers does not improve the quality of parts, $C_i = aG_i$; C is the main manufacturer's penalty for the supplier if neither improves the quality of parts, $C = a(G_1 + G_2)$. According to the above assumptions, payment matrices between suppliers are constructed, and strategy evolution of suppliers is analyzed. Payment matrices are shown in Table 1.

Table 1. Payment matrices of supplier A and supplier B.

Strategies of supplier A	Strategies of supplier B	
	Improving the quality of parts (Y)	Not improving the quality of parts ($1 - Y$)
Improving the quality of parts (X)	$G_A - C_{AH}, G_B - C_{BH}$	$G_A - C_{AH}, G_B - C_{BL} - C_B$
Not improving the quality of parts ($1 - X$)	$G_A - C_{AL} - C_A, G_B - C_{BH}$	$G_A - C_{AL} - C, G_B - C_{BL} - C$

3.2 Evolutionary Game and Stability Analysis Between Supplier a and Supplier B

Based on the payment matrix of supplier A in Table 1, expected benefit of supplier A improving the quality of parts is calculated in Eq. (1).

$$U_X = Y(G_A - C_{AH}) + (1 - Y)(G_A - C_{AH}) = G_A - C_{AH} \quad (1)$$

Expected benefit of supplier A not improving the quality of parts is as follows:

$$U_{1-X} = Y(G_A - C_{AL} - C_A) + (1 - Y)(G_A - C_{AL} - C) = G_A - C_{AL} - C + Y(C - C_A) \quad (2)$$

Average expected benefit of supplier A is as follows:

$$\begin{aligned} U_{X,1-X} &= XU_X + (1 - X)U_{1-X} \\ &= X(G_A - C_{AH}) + (1 - X)(G_A - C_{AL} - C + Y(C - C_A)) \\ &= G_A - C_{AL} - C + Y(C - C_A) + X[(G_{AL} - G_{AH} + C) - Y(C - C_A)] \end{aligned} \quad (3)$$

Expected benefit of supplier B improving the quality of parts is as follows:

$$U_Y = X(G_B - C_{BH}) + (1 - X)(G_B - C_{BH}) = G_B - C_{BH} \quad (4)$$

Expected benefit of supplier B not improving the quality of parts is as follows:

$$U_{1-Y} = X(G_B - C_{BL} - C_B) + (1 - X)(G_B - C_{BL} - C) = G_B - C_{BL} - C + X(C - C_B) \quad (5)$$

Average expected benefit of supplier B is as follows:

$$\begin{aligned} U_{Y,1-Y} &= YU_Y + (1 - Y)U_{1-Y} = Y(G_B - C_{BH}) + (1 - Y)[G_B - C_{BL} - C + X(C - C_B)] \\ &= G_B - C_{BL} - C + X(C - C_B) + Y[(C_{BL} - C_{BH} + C) - X(C - C_B)] \end{aligned} \quad (6)$$

Replicator dynamic equation of supplier A is as follows:

$$\begin{aligned} f(X) &= \frac{dX}{dt} = X(U_X - U_{X,1-X}) = X(1 - X)(U_X - U_{1-X}) \\ &= X(1 - X)[C_{AL} - C_{AH} + C - Y(C - C_A)] \end{aligned} \quad (7)$$

Replicator dynamic equation of supplier B is as follows:

$$\begin{aligned} f(Y) &= \frac{dY}{dt} = Y(U_Y - U_{Y,1-Y}) = Y(1 - Y)(U_Y - U_{1-Y}) \\ &= Y(1 - Y)[C_{BL} - C_{BH} + C - X(C - C_B)] \end{aligned} \quad (8)$$

When $\frac{dX}{dt} = 0$, $\frac{dY}{dt} = 0$, five equilibrium points of the dynamic system are solved: $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, $(x_0^*, y_0^*) = \left(\frac{C_{BL}-C_{BH}+C}{C-C_B}, \frac{C_{AL}-C_{AH}+C}{C-C_A}\right)$. X and Y represent the probability that suppliers A and B improve the quality of parts respectively, so Eq. (9) is met.

$$\begin{cases} 0 < \frac{C_{AL}-C_{AH}+C}{C-C_A} < 1 \\ 0 < \frac{C_{BL}-C_{BH}+C}{C-C_B} < 1 \end{cases} \quad (9)$$

Equation (10) is obtained by Eq. (9).

$$\begin{cases} 0 < C_{AL} - C_{AH} + C < C - C_A \\ C - C_A < C_{AL} - C_{AH} + C < 0 \end{cases} \cup \begin{cases} 0 < C_{BL} - C_{BH} + C < C - C_B \\ C - C_B < C_{BL} - C_{BH} + C < 0 \end{cases} \quad (10)$$

According to the stability of nonlinear differential equations, the system' stability is analyzed through solving the trace and determinant of Jacobian matrix of differential equations [14]. Based on stability criterion of Jacobian matrix, if the equilibrium point meets the condition: $\text{DetJ} > 0$, $\text{TrJ} < 0$, the equilibrium point is stable [15]. Jacobian matrix of $f(X), f(Y)$ is calculated as follows:

$$\begin{aligned} J &= \begin{bmatrix} \frac{df(X)}{dX} & \frac{df(X)}{dY} \\ \frac{df(Y)}{dX} & \frac{df(Y)}{dY} \end{bmatrix} \\ &= \begin{bmatrix} (1 - 2X)[C_{AL} - C_{AH} + C - Y(C - C_A)] & X(1 - X)(C_A - C) \\ Y(1 - Y)(C_B - C) & (1 - 2Y)[C_{BL} - C_{BH} + C - X(C - C_B)] \end{bmatrix} \end{aligned} \quad (11)$$

As shown in Table 2, the determinant and trace of Jacobian matrix at each equilibrium point is calculated through Eq. (11).

Table 2. Jacobian determinant and trace of each equilibrium point.

Equilibrium	Determinant	Trace
(0, 0)	$(C - C_{AH} + C_{AL})(C - C_{BH} + C_{BL})$	$2C - C_{AH} - C_{BH} + C_{AL} + C_{BL}$
(0, 1)	$(C_{AH} - C_{AL} - C_A)(C - C_{BH} + C_{BL})$	$C_A - C - C_{AH} + C_{BH} + C_{AL} - C_{BL}$
(1, 0)	$(C_{BH} - C_{BL} - C_B)(C - C_{AH} + C_{AL})$	$C_B - C + C_{AH} - C_{BH} - C_{AL} + C_{BL}$
(1, 1)	$(C_A - C_{AH} + C_{AL})(C_B - C_{BH} + C_{BL})$	$C_{AH} + C_{BH} - C_{AL} - C_{BL} - C_A - C_B$
(\bar{x}_0, \bar{y}_0)		
$\begin{pmatrix} \left(\frac{2(C - C_{AH} + C_{AL})}{C - C_A} - 1\right) \\ \left(C - C_{AH} + C_{AL} - \frac{(C - C_A)(C - C_{BH} + C_{BL})}{C - C_B}\right) \\ \left(\frac{2(C - C_{BH} + C_{BL})}{C - C_B} - 1\right) \\ \left(C - C_{BH} + C_{BL} - \frac{(C - C_B)(C - C_{AH} + C_{AL})}{C - C_A}\right) \\ -\left(\frac{C - C_{AH} + C_{AL}}{C - C_A} - 1\right) \\ (C - C_{AH} + C_{AL})\left(\left(\frac{C - C_{BH} + C_{BL}}{C - C_B} - 1\right)(C - C_{BH} + C_{BL})\right) \end{pmatrix}$		

Based on Eq. (10), there are four cases, and the stability of each equilibrium point is analyzed in different cases as shown in Table 3.

Table 3. The stability of equilibrium points in different cases.

Case	Case1 ¹			Case2 ²		
Equilibrium points	DetJ	TrJ	stability	DetJ	TrJ	stability
(0, 0)	+	+	unstable	-	uncertain	uncertain
(0, 1)	+	-	stable	-	uncertain	uncertain
(1, 0)	+	-	stable	-	uncertain	uncertain
(1, 1)	+	+	unstable	-	uncertain	uncertain
(x_0^*, y_0^*)	uncertain	uncertain	uncertain	uncertain	uncertain	uncertain
Case	Case3 ³			Case4 ⁴		
Equilibrium points	DetJ	TrJ	stability	DetJ	TrJ	stability
(0, 0)	-	uncertain	uncertain	+	-	stable
(0, 1)	-	uncertain	uncertain	+	+	unstable
(1, 0)	-	uncertain	uncertain	+	+	unstable
(1, 1)	-	uncertain	uncertain	+	-	stable
(x_0^*, y_0^*)	uncertain	uncertain	uncertain	uncertain	uncertain	uncertain

¹Case1: $0 < C_{AL} - C_{AH} + C < C - C_A \cup C_{BL} - C_{BH} + C < C - C_B$

²Case2: $C - C_A < C_{AL} - C_{AH} + C < 0 \cup 0 < C_{BL} - C_{BH} + C < C - C_B$

³Case3: $0 < C_{AL} - C_{AH} + C < C - C_A \cup C_B - C_{BL} - C_{BH} + C < 0$

⁴Case4: $C - C_A < C_{AL} - C_{AH} + C < 0 \cup C - C_B < C_{BL} - C_{BH} + C < 0$

4 Simulation Analysis

According to stability conditions of equilibrium points in Table 3, the initial values of payment matrix parameters are set: $G_A = 100$ ten thousand dollars, $G_B = 90$ ten thousand dollars, $C_{AH} = 92$ ten thousand dollars, $C_{AL} = 12$ ten thousand dollars, $C_{BH} = 88$ ten thousand dollars, $C_{BL} = 10$ ten thousand dollars.

Suppliers dynamically change strategies through continuous comparison and learning. The penalty of the main manufacturer for not improving the quality of parts plays an important role in the strategy evolution of suppliers, and the impact of penalty coefficient on the strategy evolution of suppliers is analyzed.

Supplier A and supplier B randomly select the initial probability, and the initial probability of improving quality of parts is set: (0.5,0.5).

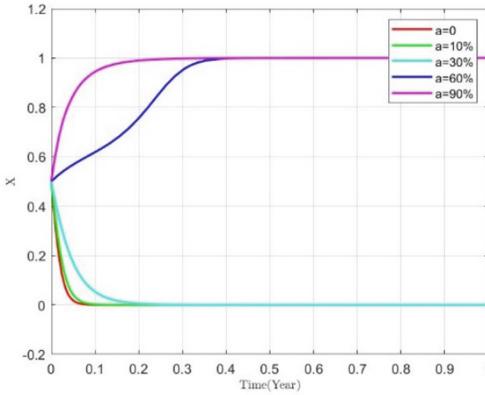


Fig. 1. The impact of penalty coefficient on strategy evolution of supplier A of (0.5,0.5).

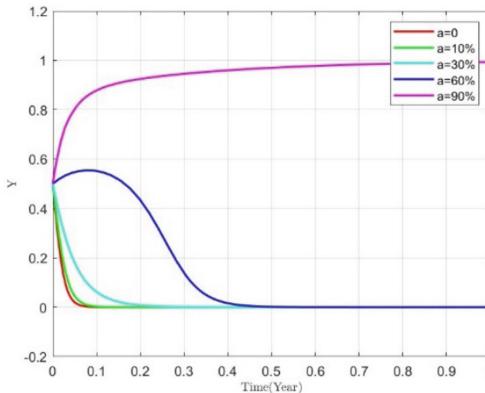


Fig. 2. The impact of penalty coefficient on strategy evolution of supplier B of (0.5,0.5).

Figure 1 and Fig. 2 show the impacts of penalty coefficient on strategy evolution of suppliers. In Fig. 1, if penalty coefficient is not more than 30%, supplier A stabilizes to equilibrium state “0”, and supplier A does not improve the quality of parts. As the penalty coefficient decreases, evolution speed to the equilibrium state “0” is faster. If penalty coefficient is not less than 60%, supplier A stabilizes to equilibrium state “1”, and supplier A improves the quality of parts. Evolution time of stabilizing to the equilibrium state “1” is shorter with penalty coefficient increasing. The evolution analysis of supplier B is similar to that of supplier A. The difference is that penalty coefficient is not more than 60% when supplier B stabilizes to equilibrium state “0”.

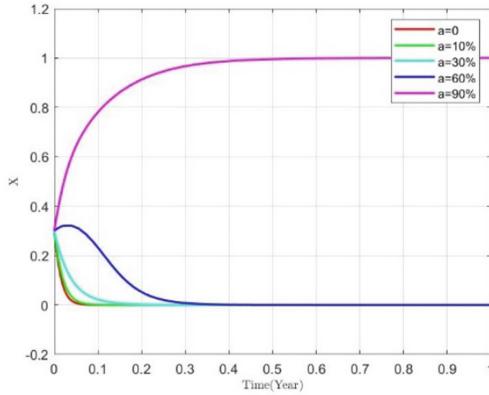


Fig. 3. The impact of penalty coefficient on strategy evolution of supplier A of (0.3,0.5).

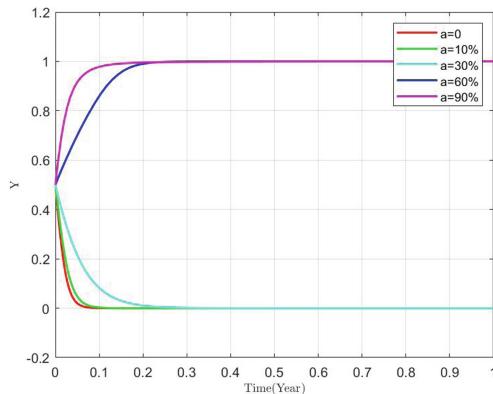


Fig. 4. The impact of penalty coefficient on strategy evolution of supplier B of (0.3,0.5).

The initial probability of supplier A and supplier B is set: (0.3,0.5). As penalty coefficient increases from 0 to 90%, evolutionary trajectories of suppliers' strategies are illustrated in Fig. 3 and Fig. 4. In Fig. 3, with the increase of penalty coefficient, the time to reach a stable state "0" is longer, indicating that supplier A tends to choose the NI strategy. In Fig. 4, the higher penalty coefficient from 0 to 30%, the slower the convergence rate to the equilibrium state "0". Supplier B stabilizes to the equilibrium state "1" with penalty coefficient increasing from 60% to 90%, and supplier B tends to improve the quality of parts. Therefore, the increase in the value of penalty coefficient enhances suppliers' enthusiasm for choosing the I strategy.

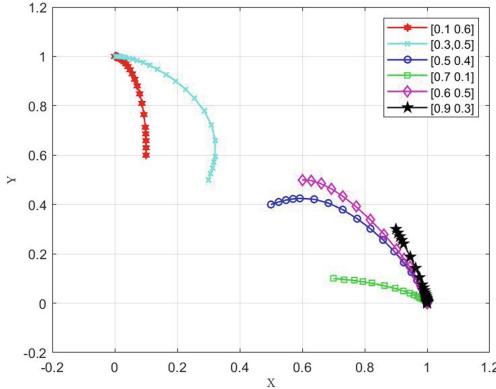


Fig. 5. The impact of initial probability on strategy evolution of suppliers.

Figure 5 illustrates the impact of initial probability on strategy evolution of suppliers. When the initial probability is (0.1,0.6) and (0.3,0.5) respectively, the evolution finally converges to the stable state:(not improve the quality of parts, improve the quality of parts). When the initial probability is (0.5,0.4), (0.7,0.1), (0.6,0.5) and (0.9,0.3) respectively, the system converges to the stable point (1, 0), and the equilibrium strategy is (improve the quality of parts, not improve the quality of parts).

5 Conclusion

This paper establishes a two-party game model between supplier A and supplier B, and analyzes the impact of penalty coefficient of the main manufacturer on the strategy evolution of suppliers through numerical simulation. The result shows that:

As penalty coefficient of the main manufacturer increases, suppliers tend to improve the quality of parts. In order to effectively improve the quality of parts, the main manufacturer should charge a higher fine on suppliers who do not improve the quality of parts to restrain the suppliers' behaviors.

In the future, the game model is verified based on real data of enterprises. Additionally, this paper establishes an evolutionary game between two suppliers, but the main manufacturer plays an important role in the manufacturing process. The evolutionary game model between the main manufacturer and the supplier can be considered in the future.

Acknowledgements. The work is supported by the national natural science foundation of China under Grant No.72171173; Shanghai Science and Technology Innovation Action Plan under Grant No.19DZ1206800.

References

1. Zhang, M., Zhu, J.J., Wang, H.H., Liu, P.: Evolutionary game analysis on strategies in “main manufacturer–supplier” mode considering technology docking and price concluding under competition condition. *Mathematics* **7**(12), 1–25 (2019)

2. Wang, X.H.: Pre-sale pricing strategies based on the reputation of an e-commerce platform enterprise and herd effect. *J. Industrial Eng. Eng. Manage.* **35**(1), 126–141 (2021)
3. Weng, Z.C., Luo, P.L.: Supervision of the default risk of online car-hailing platform from an evolutionary game perspective. *Sustainability* **13**(2), 1–19 (2021)
4. Li, J.Q., Ren, H., Zheng, J.: Stochastic evolutionary game analysis of multiple supervision paths of enterprise R&D manipulation. *Chinese J. Manage. Sci.* **29**(10), 191–201 (2021)
5. Cheng, Y., Bi, L.N., Tao, F., Ji, P.: Hypernetwork-based manufacturing service scheduling for distributed and collaborative manufacturing operations towards smart manufacturing. *J. Intell. Manuf.* **31**(7), 1707–1720 (2018)
6. Shan, Z.D., Chen, L., Han, L.L., Zeng, Y.H.: Decision mechanism of multi-participators service innovation behavior in digital manufacturing environment. *Computer integrated manufacturing system*, 1–24 (2022)
7. Yang, K., Wang, W., Hu, B.: Evolutionary game models on multiagent collaborative mechanism in responsible innovation. *Sci. Program.* **2020**, 1–11 (2020)
8. Wang, M.M., Lian, S., Yin, S., Dong, H.M.: A three-player game model for promoting the diffusion of green technology in manufacturing enterprises from the perspective of supply and demand. *Mathematics* **8**(9), 1–26 (2020)
9. He, J.J., Jiang, X.L., Xu, F.Y.: Analysis of evolution path of cooperation game model based on SDN enterprises. *Operations Res. Manage.* **27**(09), 83–90 (2018)
10. Li, Q., Kang, Y.F., Tan, L.L., Chen, B.: Modeling formation and operation of collaborative green innovation between manufacturer and supplier: a game theory approach. *Sustainability* **12**(6), 2209 (2020)
11. Zhang, S.Z., Yu, Y.M., Zhu, Q.H., Qiu, C.M., Tian, A.X.: Green innovation mode under carbon tax and innovation subsidy: an evolutionary game analysis for portfolio policies. *Sustainability* **12**(4), 1385 (2020)
12. Li, Z., Jin, G.H., Duan, S.: Evolutionary game dynamics for financial risk decision-making in global supply chain. *Complexity* **2018** (2018)
13. Ma, J.H., Lou, W.D., Tian, Y.: Bullwhip effect and complexity analysis in a multi-channel supply chain considering price game with discount sensitivity. *Int. J. Prod. Res.* **57**(17), 5432–5452 (2019)
14. Gao, X.R., Shi, Y., Chen, Z.L.: Evolutionary game research on credit strategy of sharing economic transaction subjects. *J. Chongqing Normal Univ.: Natural Science Ed.* **37**(1), 10 (2020)
15. Tong, W., Mu, D., Zhao, F., Gamini, P.M., John, W.S.: The impact of cap-and-trade mechanism and consumers' environmental preferences on a retailer-led supply chain. *Resour. Conserv. Recycl.* **142**, 88–100 (2019)



Multi-party Evolution Stability Analysis of Electric Vehicles- Microgrid Interaction Mechanism

Haitong Guo, Hao Zhang^(✉), Jianfeng Lu, Rong Zeng, and Tiaojuan Han

CIMS Research Center, Tongji University, Shanghai 201804, China
hzhang@tongji.edu.cn

Abstract. In the process of interaction between the electric vehicle (EV) and the microgrid (MG), the discharge electricity price provided by the microgrid is a key factor affecting the benefits of the participants. This paper uses evolutionary game theory combined with system dynamics to simulate the dynamic process of multi-parties game in the process of EV-MG interaction under the condition of bounded rationality and analyzes the influence of different discharge price pricing strategy on the game process. First of all, when the formulating strategy of the discharge price is static, the evolutionary strategy of the participants will tend to be stable, but make strategy stabilization delay time longer, which will increase the extra game cost of the participants. Secondly, when the strategy for formulating discharge price is dynamic, it can not only make the strategy of the participants stable, but also have the convergence speed of the game process less affected by the initial value, which can significantly reduce the game cost of all parties. Combining system dynamics and evolutionary game theory to study the interaction process of EV-MG provides an effective solution for microgrid to formulate electric vehicle discharge price strategy.

Keywords: Electric vehicle · Microgrid · Interaction · Evolutionary game · System dynamics · Pricing strategy

1 Introduction

The new energy microgrid is an important development trend and direction of the future energy market [1]. Compared with traditional energy sources, new energy microgrids are more in line with the concept of sustainable development [2]. However, the power generation capacity of new energy is easily affected by the environment and climate, and has intermittency and randomness, which adds complexity to the operation management and control of the microgrid [3]. In addition, electric vehicles as a new demand for green consumption have been rapidly promoted and popularized in recent years. This means that a large number of electric vehicles can become important mobile energy storage units, which can help the microgrid solve intermittent problems and reduce operating costs after reasonable scheduling [4]. The new energy microgrid can also reduce the

dependence of electric vehicles on fossil fuel power generation and the use cost of electric vehicles [5], and realize low carbon in the true sense.

Actually, in the interaction between EVs and new energy MGs, their benefits are both opposed and interdependent. The interaction between them is game-like. In the existing literature on EV-MG interaction, game theory is widely used to analyze conflicts of interest. References [6, 7] used game theory to establish a planning model with the goal of minimizing grid operating costs and EV users, and obtained Nash equilibrium solutions for both. However, the above research is only a static analysis of the game between EV and the MG, ignoring the dynamic process of the game. References [8, 9] use evolutionary game theory to establish a game model, and obtain a discharge price that meets the interests of both parties. But they just did research on one side in the way of the static pricing strategies, and neither considered the different benefit needs of different types of EVs nor the practicality of static pricing strategies. References [10, 11] just analyzed the influence of electricity price on the change of microgrid strategy. These studies provided useful ideas for the formulation of discharge prices. However, in the process of EV-MG interaction, participants' strategies are not always static. When there are different types of EVs, the interests and requirements between EVs are also different. They dynamically change strategies by observing and comparing their benefits with others and adjusting their strategy choices, which requires the study of dynamic games involving multiple parties under bounded rationality. Therefore, this paper combines evolutionary game theory and system dynamics to describe the dynamic process of the multi-party game of EV-MG interaction and analyzes the impact of different pricing strategies on the game process and game equilibrium.

The rest of this paper is organized as follows: Sect. 2 introduces the evolutionary game process of the EV-MG interaction. Section 3 introduces the multi-party evolutionary game simulation of EV-MG interaction based on system dynamics. Section 4 discusses the comparison of static and dynamic pricing strategy of MG for EV discharge price. Finally, the concluding observations are published in Sect. 5.

2 Multi-party Evolutionary Game of EV-MG Interaction

Evolutionary game theory is a theory that combines traditional game theory with dynamic evolutionary analysis. Unlike traditional game theory, evolutionary game does not require both parties involved in the game to be completely rational, nor to obtain complete information [12, 13]. Evolutionary game emphasizes dynamic equilibrium. It analyzes the possible income of each party and analyzes the income trend of each party under different strategy, so as to obtain the evolutionary stable strategy of the game parties under different circumstances. In the process of EV-MG interaction, the participants are bounded rational, and there are different interests and requirements between different types of EVs. Therefore, evolutionary game theory is suitable for studying the dynamic game of bounded rational participants in the process of EV-MG interaction.

3 Designing the Game Model

Assumption 1:Brand A EV car, brand B EV car(short for EV A,EV B) and microgrid are bounded rational and aim to maximize their own interests.

Assumption 2:The three are in a state of information asymmetry. In the game process, the microgrid is in a dominant position, and the EV A and the EV B are in a subordinate position.

Assumption 3: The rate of EV A choosing cooperation is $x(0 \leq x \leq 1)$, the rate of choosing not to coopeate is $1 - x$; the rate of EV B choosing cooperation is $y(0 \leq y \leq 1)$, The rate of choosing non-cooperation is $1 - y$; the rate of microgrid choosing cooperation $z(0 \leq z \leq 1)$, and the rate of choosing non-cooperation is $1 - z$.

Assumption 4:When the two kinds of EV choose the non-cooperative strategy, they will only charge in the microgrid. The charging electricity of the EV A is EQ_{c1} and the charging electricity of the EV B is EQ_{c2} . The charging price is P_{c1} and P_{c2} respectively. When EV A and EV B choose the cooperative strategy, besides charging in the microgrid, they will also discharge to the microgrid. The amount of electricity discharged by the EV A is EQ_{dis1} , and the discharged electricity of the EV B is EQ_{dis2} . If the microgrid does not cooperate at this time, they respectively have a part of the loss L_{u1} and L_{u2} . When the two vehicles choose the cooperation independently, the discharge price is P_{dis1} , P_{dis2} , and when both choose the cooperation,the discharging price is P_{dis} .

Assumption 5:The normal income of the microgrid is R_{mg} . When the microgrid chooses to cooperate, it needs to spend extra on infrastructure improvement and the unit cost is Ex .but the reverse discharge of EV will balance part of the electricity from the external grid during peak power consumption. If EV A does not cooperate at this time, the microgrid will lose L_1 . If EV B does not cooperate at this time, the microgrid will lose L_2 . If both vehicles do not cooperate, the loss will be L .

The above-mentioned EV-MG interaction multi-party game and its corresponding variables are shown in Fig. 1 and Table 1 below. Therefore, according to the previous assumptions and analysis, the payoff matrix of the multi-parties game can be obtained as shown in Table 2.

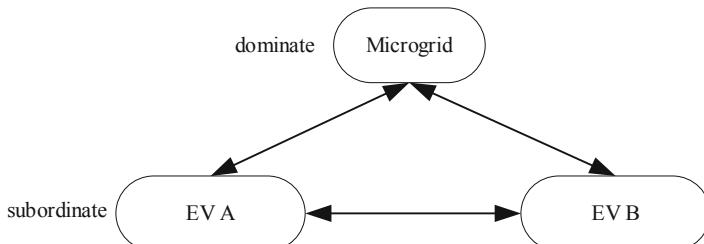


Fig. 1. Multi-players game of EV-microgrid interaction.

Table 1. Meanings of the variables in the multi-players game.

Variables	Meanings of the variables	Notes
P_{dis}	discharge price of tripartite cooperation	$P_{dis} > 0$
P_{disi}	Discharge price for only single brand EV cooperation and microgrid cooperation	$P_{disi} > 0 \text{ and } i = 1, 2$
P_{ci}	EV A charging price	$P_{ci} > 0 \text{ and } i = 1, 2$
L_{bi}	Unit battery loss of EV	$L_{bi} > 0 \text{ and } i = 1, 2$
EQ_{disi}	Discharge capacity of EV	$EQ_{disi} > 0 \text{ and } i = 1, 2$
EQ_{ci}	The charging capacity of the EV	$EQ_{ci} > 0 \text{ and } i = 1, 2$
L_{ui}	Single Brand EV cooperation loss that does not cooperate with microgrid	$L_{ui} > 0 \text{ and } i = 1, 2$
L_i	The EV loss of microgrid cooperation but EV not cooperation	$L_i > 0 \text{ and } i = 1, 2$
R_{mg}	Normal benefits of microgrids	$R_{mg} > 0$
Ex	Additional unit costs for microgrid infrastructure	$Ex > 0$
P_g	Electricity purchase price from external grid	$P_g > 0$
L	Microgrid losses for microgrid cooperation but two types of EV not cooperation	$L > 0$
x	EV A cooperation rate	$0 \leq x \leq 1$
y	EV B cooperation rate	$0 \leq y \leq 1$
z	Microgrid cooperation rate	$0 \leq z \leq 1$

Table 2. The payoff matrix of the multi-parties game

EV A	EV B	Microgrid	
		Cooperative strategy	Non-cooperative strategy
Cooperative strategy	Cooperative strategy	$P_{dis} * EQ_{dis1} - P_{c1} * EQ_{c1} - (2EQ_{dis1} + EQ_{c1}) * L_{b1}$	$-P_{c1} * EQ_{c1} - EQ_{c1} * L_{b1} - L_{u1}$
		$P_{dis} * EQ_{dis2} - P_{c2} * EQ_{c2} - (2EQ_{dis2} + EQ_{c2}) * L_{b2}$	$-P_{c1} * EQ_{c1} - EQ_{c1} * L_{b1} - L_{u2}$
		$R_{mg} + P_{c1} * EQ_{c1} + P_{c2} * EQ_{c2} + (P_g - P_{dis1} - P_{c1}) * EQ_{dis1} + (P_g - P_{dis2} - P_{c2}) * EQ_{dis2} - Ex$	$R_{mg} + P_{c1} * EQ_{c1} + P_{c2} * EQ_{c2}$
	Non-cooperative strategy	$P_{dis1} * EQ_{dis1} - P_{c1} * EQ_{c1} - (2EQ_{dis1} + EQ_{c1}) * L_{b1}$	$-P_{c1} * EQ_{c1} - EQ_{c1} * L_{b1} - L_{u1}$
		$-P_{c2} * EQ_{c2} - EQ_{c2} * L_{b2}$	$-P_{c2} * EQ_{c2} - EQ_{c2} * L_{b2}$
		$R_{mg} + P_{c1} * EQ_{c1} + P_{c2} * EQ_{c2} + (P_g - P_{dis1} - P_{c1}) * EQ_{dis1} - Ex - L_2$	$R_{mg} + P_{c1} * EQ_{c1} + P_{c2} * EQ_{c2}$
Non-cooperative strategy	Cooperative strategy	$-P_{c1} * EQ_{c1} - EQ_{c1} * L_{b1}$	$-P_{c1} * EQ_{c1} - EQ_{c1} * L_{b1}$

(continued)

Table 2. (continued)

EV A	EV B	Microgrid
		Cooperative strategy
Non-cooperative strategy		$P_{dis2}^2 * EQ_{dis2} - P_{c2}^2 * EQ_{c2} - (2EQ_{dis2} + EQ_{c2}) * L_{b2}$
		$R_{mg} + P_{c1} * EQ_{c1} + P_{c2} * EQ_{c2} + (P_g - P_{dis2} - P_{c2}) * EQ_{dis2} - Ex - L_1$
		$-P_{c1} * EQ_{c1} - EQ_{c1} * L_{b1}$
		$-P_{c2} * EQ_{c2} - EQ_{c2} * L_{b2}$
		$R_{mg} + P_{c1} * EQ_{c1} + P_{c2} * EQ_{c2} - Ex - L$
		Non-cooperative strategy

3.1 Describing the Game Strategy

According to evolutionary game theory, the replication dynamic equation is used to represent the learning and evolution process of participants in the process of EV-MG interaction. Therefore, the cooperative strategy expected benefits U_x of EV A and the non-cooperative strategy expected benefits U_{1-x} , the cooperative strategy expected benefits U_y of EV B and the non-cooperative strategy expected benefits U_{1-y} and the cooperative strategy expected benefits U_z of the microgrid and the non-cooperative strategy expected benefits U_{1-z} can be respectively obtained from the following ways.

3.1.1 The Game Strategy of EV A

For the EV A, the benefits of choosing cooperation are composed of three parts: discharge revenue, charging cost and battery loss. and the expected cooperation income after simplification is:

$$U_x = yz(P_{dis} - P_{dis1})EQ_{dis1} + z(P_{dis1}EQ_{dis1} - 2EQ_{dis1}L_{b1} + L_{u1}) + (-P_{c1}EQ_{c1} - EQ_{c1}L_{b1} - L_{u1}) \quad (1)$$

The expected non-cooperation income after simplification is:

$$U_{1-x} = -P_{c1}EQ_{c1} - EQ_{c1}L_{b1} \quad (2)$$

The replicate dynamic equation is:

$$F(x) = \frac{dx}{dt} = x(1-x)(U_x - U_{1-x}) = x(1-x)(yz(P_{dis} - P_{dis1})EQ_{dis1} + z(P_{dis1}EQ_{dis1} - 2EQ_{dis1}L_{b1} + L_{u1}) - L_{u1}) \quad (3)$$

The game strategy of EV B.

Similarly, the expected cooperation income U_y and the expected non-cooperation income U_{1-y} after simplification are:

$$U_y = xz(P_{dis} - P_{dis2})EQ_{dis2} + z(P_{dis2}EQ_{dis2} - 2EQ_{dis2}L_{b2} + L_{u2}) + (-P_{c2}EQ_{c2} - EQ_{c2}L_{b2} - L_{u2}) \quad (4)$$

$$U_{1-y} = -P_{c2}EQ_{c2} - EQ_{c2}L_{b2} \quad (5)$$

The replicate dynamic equation is:

$$\begin{aligned} F(y) = \frac{dy}{dt} &= y(1-y)(U_y - U_{1-y}) = y(1-y)(xz(P_{dis} - P_{dis2})EQ_{dis2} \\ &+ z(P_{dis2}EQ_{dis2} - 2EQ_{dis2}L_{b2} + L_{u2}) - L_{u2}) \end{aligned} \quad (6)$$

The game strategy of microgrid.

For the microgrid, the benefits of the cooperation strategy consist of normal benefits, EV A charging benefits, EV B charging benefits, peak shaving and valley filling benefits, and input costs. The non-cooperative strategy is composed of normal benefits, EV A charging revenue and EV B charging revenue. Therefore, the expected benefits of the cooperation and non-cooperation strategy of the microgrid after simplification is

$$\begin{aligned} U_z &= xy * ((P_{dis1} - P_{dis})EQ_{dis1} + (P_{dis2} - P_{dis})EQ_{dis2}) + x((P_g - P_{dis1} - P_{c1})EQ_{dis1} \\ &+ L - L_2) + y((P_g - P_{dis2} - P_{c2})EQ_{dis2} + L - L_1) + (R_{mg} + P_{c1}EQ_{c1} + P_{c2}EQ_{c2} - E_x - L) \end{aligned} \quad (7)$$

$$U_{1-z} = R_{mg} + P_{c1}EQ_{c1} + P_{c2}EQ_{c2} \quad (8)$$

The replicate dynamic equation is:

$$\begin{aligned} F(z) &= z(1-z)(U_z - U_{1-z}) = z(1-z)(xy(P_{dis1} - P_{dis})EQ_{dis1} + (P_{dis2} - P_{dis})EQ_{dis2} \\ &+ x((P_g - P_{dis1} - P_{c1})EQ_{dis1} + L - L_2) + y((P_g - P_{dis2} - P_{c2})EQ_{dis2} + L - L_1) \\ &- E_x - L) \end{aligned} \quad (9)$$

3.2 Game Solution Analysis

From the analysis of 2.2, the replication dynamic equations of the three-party game are shown in Eq. 3, Eq. 6 and Eq. 9. In the actual process of EV-MG interaction, electric vehicles of different brands are not completely independent, but exist in various interest relationships. Different electric vehicles also have different interests and requirements on whether to choose a cooperative strategy. In addition, with the increase of players participating in the game, the benefit relationship caused by different strategies becomes more and more complicated. In this case, it is much more difficult to analyze the stability of the evolutionary game of EV-MG interaction. It is obvious to see from the equations that it is not easy to find all equilibrium points of multiple players and analyze their stability. Therefore, when the model analysis cannot fully utilize the theoretical analysis to achieve the original purpose, the simulation method can be used to simulate the implementation effect of different strategies and make scientific predictions [14, 15].

3.3 Building the Simulation Model Based on System Dynamics

The biggest advantage of system dynamics (SD) is to focus on the dynamic changes of the system and the influence of parameter changes on the results, which can solve

various complex problems [13]. In order to better analyze and compare the interests and needs of different parties, this section will establish a multi-player game model based on SD. Using Vensim PLE 7.3.5, according to the analysis assumptions in Sect. 2, a multi-parties evolutionary game model of EV-MG interaction is established. The participants in this model are microgrid, EV A and EV B.

The model set initial time = 0, final time = 8, time step = 0.01, time unit: days, integration type: Euler. According to the relevant data of the Shanghai Electric Power University Lingang microgrid experimental platform, the initial values of each auxiliary variable of the system dynamics model are preprocessed as shown in the Table 3.

Table 3. Initial values of auxiliary variables in the system dynamics model.

Variables	Meanings of the variables	Values
P_{ci}	EVs charging price	$P_{c1} = 0.5, P_{c2} = 0.4$
L_{bi}	Unit battery loss of EV	$L_{b1} = 0.05, L_{b2} = 0.06$
EQ_{disi}	Discharge capacity of EV	$EQ_{dis1} = 30, EQ_{dis2} = 15$
EQ_{ci}	The charging capacity of the EVs	$EQ_{c1} = 60, EQ_{c2} = 30$
L_{ui}	Single Brand EV cooperation loss that does not cooperate with microgrid	$L_{u1} = 2, L_{u2} = 1$
L_i	The EV loss of microgrid cooperation but EV not cooperation	$L_1 = 3, L_2 = 2$
R_{mg}	Normal benefits of microgrids	$R_{mg} = 1000$
Ex	Additional unit costs for microgrid infrastructure	$Ex = 1$
P_g	Electricity purchase price from external grid	$P_g = 0.9$
L	Microgrid losses for microgrid cooperation but two types of EV not cooperation	$L = 5$
x	Brand A EV cooperation rate	0.5
y	Brand B EV cooperation rate	0.5
z	Microgrid cooperation rate	0.5

3.3.1 Static Pricing Strategies

In the process of EV-MG interaction, it is a common control strategies for the microgrid to use a fixed price as a means of payment for the discharge behavior of EVs. Therefore, in order to study the effect of different fixed discharge prices on the enthusiasm of electric vehicles to participate in the EV-MG interaction, this section will simulate the effect of different fixed discharge prices.

When the pricing strategies of the microgrid is a static pricing strategies, the income of the corresponding strategies is constant. In addition, the participants are rational, assuming that the rate of their initial cooperation is 0.5. After the evolutionary game,

they will dynamically change their own strategies and then adjust strategies. In the EV-MG interaction evolutionary game dynamics model, the static pricing changes from 0.1 to 0.3 and 0.5. The game results of each party are shown in the figures.

In Fig. 2, the red curve, the blue curve, and the green curve respectively represent the changes in the cooperation rate x of the EV A with a fixed price of 0.1, 0.3 and 0.5 under the static discharge strategies. Correspondingly, Fig. 3 and the Fig. 4 respectively represent the cooperation rate of the EV B and the microgrid under the same conditions. It can be seen from the Fig. 2 and Fig. 3 that under the same discharge electricity price, because the charge and discharge capacity of EV B is smaller than that of EV A, the trend change is slower than that of EV As. Although there are differences in charging price and charging and discharging capacity between EV A and EV B, the trend of change is roughly the same. As can be seen from the above two figures, when the fixed discharge price is 0.1, the cooperation rate gradually decreases to 0 because the benefits

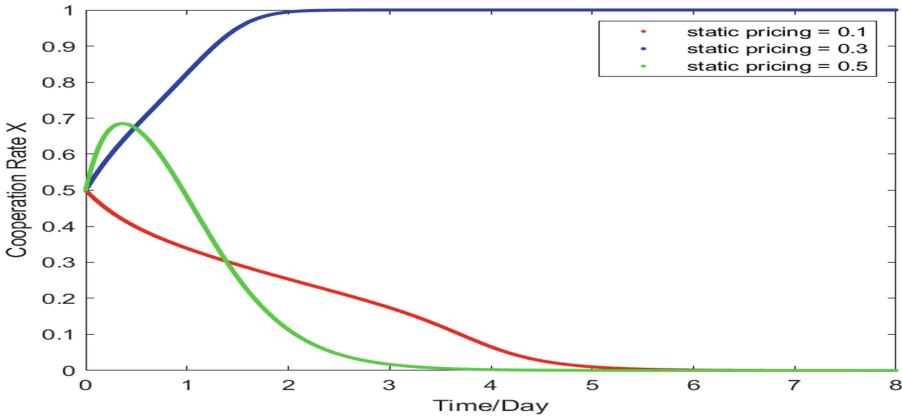


Fig. 2. The change curve of EV A cooperation rate

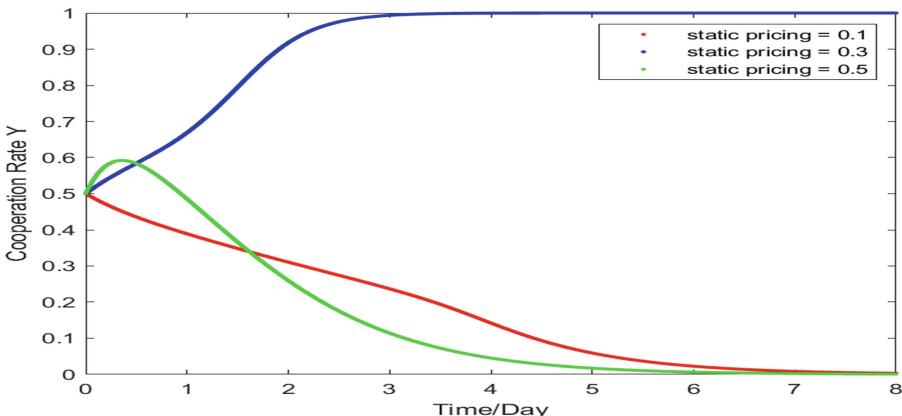


Fig. 3. The change curve of EV B cooperation rate

of electric vehicle participants are less than the benefits of non-cooperation. When the fixed price is 0.3, the benefits of electric vehicle cooperation are greater than the benefits of non-cooperation, so they will choose to cooperate. When the fixed price is 0.5, at first, because the electric vehicles can get far more benefits than non-cooperation, the cooperation rate of the electric vehicles increases. But then because of non-cooperation of microgrid, their earnings also bear the cost of risk, so they tend to be less cooperative.

Correspondingly, it can be seen from Fig. 4 that when the fixed discharge price is 0.1, the initial cooperation benefit of the microgrid is much greater than the non-cooperative benefit, so the cooperative strategies is inclined. However, due to the non-cooperation of EVs, the cost of the cooperation strategies of microgrid is greater than the benefit, so it gradually tends to the non-cooperation strategies. When the fixed price is 0.3, for the microgrid, the cooperation rate of electric vehicles was not very high at first, and the cooperation income of microgrids was slightly less than that of non-cooperation, so the cooperation rate dropped a little. However, due to the gradual increase of the vehicle cooperation rate, the benefit of the microgrid increases, which exceeds the benefit of the non-cooperation of the microgrid, so it tends to cooperate in the end. When the fixed price is 0.5, obviously, the benefit of the microgrid cooperation strategies is less than that of the non-cooperative strategies in this case, so it gradually tends to the non-cooperative strategy.

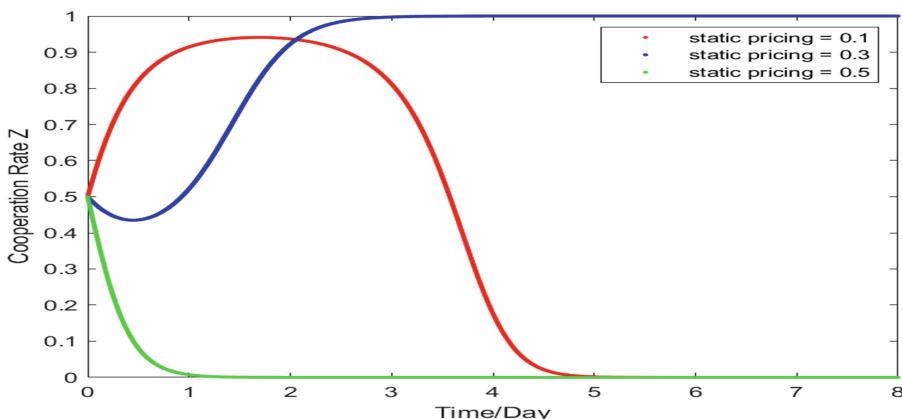


Fig. 4. The change curve of microgrid cooperation rate

The above simulation results show that at different fixed discharge prices, the strategy choices of each participant will change and fluctuate, and it also shows that there is an evolutionary and stable strategies in the process of EV-MG interaction. However, although the static fixed price can stabilize the evolution strategies, the delay time is relatively long, which will increase the extra cost of each participant.

3.3.2 Dynamic Pricing Strategies

The existing strategies make the control process delay relatively long. Therefore, it is necessary to study more efficient and stable control strategies. References [14, 15] have shown that associating policies with their corresponding behavior ratios can effectively speed up the convergence. Therefore, this paper proposes a dynamic pricing strategies, that is, the microgrid sets the discharge price according to the proportion of electric vehicles participating in the interaction between EV and MG, shortens the switching time of the strategies of all parties during the EV-MG interaction, and reduces the losses of all parties. The specific formula is as follows:

$$\begin{cases} P_{dis1} = value - value(1 - x) \\ P_{dis2} = value - value(1 - y) \\ P_{dis} = value - value(1 - x)(1 - y) \end{cases} \quad (10)$$

Thus, when the electric vehicles do not participate in the cooperation, the discharge price is 0. In the case of participating in the cooperation, as the cooperation rate increases, the maximum value can be reached. In the process of EV-MG interaction, we assume that all participants are rational, and the microgrid is the one that sets the electricity price for discharge and in a dominant position. Therefore, modeling the dynamic pricing strategy, this paper selects the initial value of 0.1 and 0.3 to compare with participant EV A in a subordinate position, whose strategies changes faster in the static pricing strategies. The simulation results are shown in Fig. 5 and Fig. 6.

The Fig. 5 and Fig. 6 respectively show the impact of different pricing strategies on the changes of EV A and microgrid cooperation rate. In the Fig. 5, the red thin line and the blue thin line respectively represent the strategy evolution process of the EV A with the initial value of 0.1 and 0.3 under the dynamic pricing strategies. Correspondingly, in the Fig. 6, the red thin line and the blue thin line represent the evolution strategy process of microgrid under the same conditions.

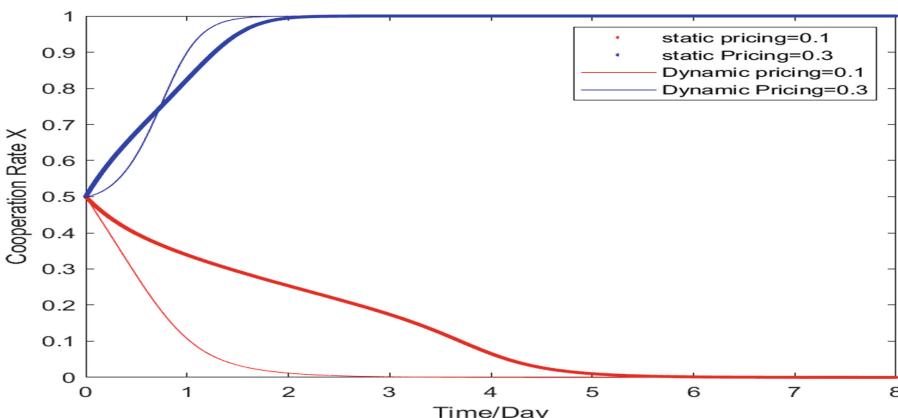


Fig. 5. Effect of different pricing strategies on EV A.

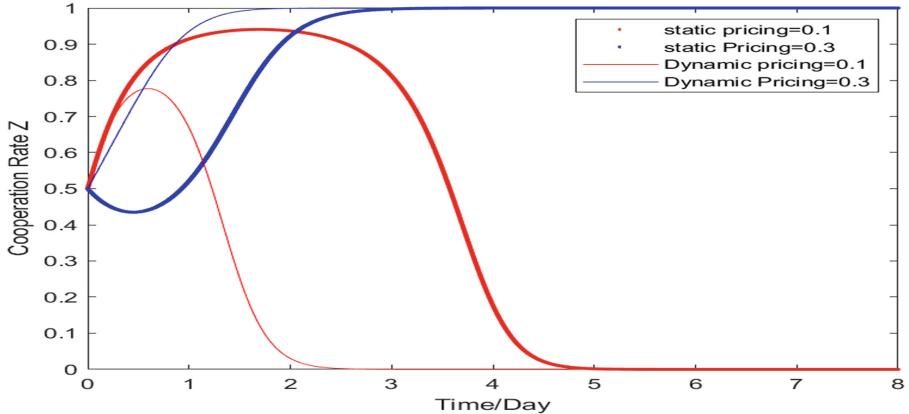


Fig. 6. Effect of different pricing strategies on Microgrid.

Comparing the two figures, when the initial value of the dynamic discharge price is equal to the fixed value of static pricing, it can be seen that the state of EV A and the state of microgrid under dynamic pricing have greatly improved the convergence speed compared with that under static pricing strategies. When the initial value of dynamic pricing is 0.3, the initial cooperation rate has not reached a high level. Therefore, the benefit of EV A is smaller than that of static pricing. However, as the cooperation rate increases, the benefit will increase, prompting the vehicle cooperation rate to converge faster. Correspondingly, in terms of microgrids, the initial lower discharge electricity price makes the benefit of the microgrid balance the input cost in the EV-MG interaction, and also makes the change of the cooperation rate of the microgrid not tend to be in the non-cooperative state, thus significantly reducing the delay time of the microgrid.

3.3.3 Comparison of Stability

In order to test the influence of static pricing strategies and dynamic pricing strategies on the evolutionary stability of EV-MG interaction when each participant chooses the initial cooperation rate. This section sets the different initial cooperation rates of evolution processes under two strategies. The initial cooperation rate (x, y, z) is set to $(0.5, 0.5, 0.5)$ and $(0.7, 0.7, 0.7)$, respectively and final time = 16. The simulation results are shown in Fig. 7. and Fig. 8.

The thin line in the figure represents the change of the cooperation rate of all parties in the dynamic pricing strategy, and the thick line represents the effect under the static pricing strategy. It can be seen from the comparison in the figures that, compared with the static pricing strategies, the delay time of the cooperation strategies transition under the dynamic pricing strategies is less affected by the initial cooperation rate. Therefore, although the adoption of static pricing strategies in the process of EV-MG interaction can ultimately make both EV-MG and EV-MG a stable strategic state, it is greatly affected by different factors, and the strategy changes takes a long time. The dynamic pricing strategies is more stable, which can effectively speed up the convergence of the strategies in the game process.

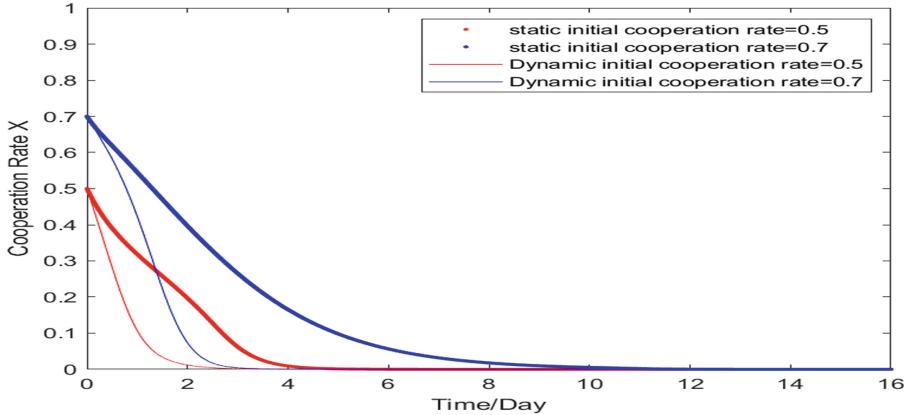


Fig. 7. Effect of different initial values on the EV A under different penalty strategies.

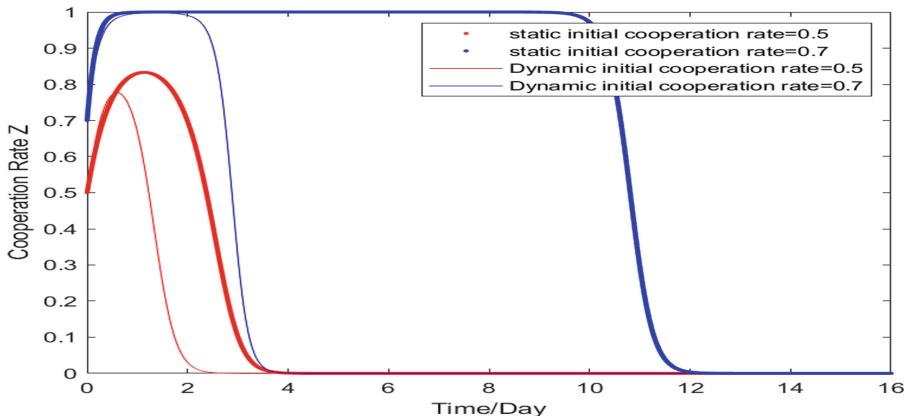


Fig. 8. Effect of different initial values on the Microgrid under different penalty strategies.

4 Discussion

When the discharge electricity price is a static pricing strategies, an appropriate increase in the discharge price will attract electric vehicles to participate in the EV-MG interaction process, and other electric vehicles will also participate in this process when they realize that it is profitable. However, this pricing method will increase the time for the participants to change their strategies and increase the game costs of the participants. When some electric vehicle users have a strong willingness to participate, and the discharge electricity price from the microgrid is relatively low, at this time, electric vehicle users are unprofitable and tend not to cooperate. However, the strategies of the microgrid is still in the state of cooperation, and the cost related to the cooperation is continuously invested. This process greatly increases the cost of the participants in the EV-MG interaction (as shown in the Fig. 7 and Fig. 8).

Existing literature studies on the electric vehicle discharge price pricing strategies will easily lead to a long evolutionary cycle of the participants' strategies. If the discharge price is arbitrarily changed, it not only will reduce the expectations of electric vehicle users for the EV-MG interaction, but also will make the interests of the grid damaged. Therefore, in order to reduce the cost of the game for the microgrid and more quickly know the willingness of electric vehicle users to the EV-MG interaction. A dynamic pricing strategy is a great option.

5 Conclusion

The EV-MG interaction is a dynamic evolution process, and a suitable electric vehicle discharge price can promote the benign interaction between MG and EVs. This paper uses evolutionary game theory to describe the dynamic process between microgrid and two different brands of EVs under bounded rationality. The main research results are as follows.

- (1) When in static pricing strategy condition, the three participants have a stable evolutionary stability strategies. However, it takes a long time for the strategies to stabilize, which will easily lead to more costs for the participants in the process of EV-MG interaction. In addition, the fluctuation range of each party's strategies and the stable time of strategies evolution are greatly affected by the initial value.
- (2) When in dynamic pricing strategy condition, the participants not only have a stable evolutionary stability strategies, but also can significantly reduce the delay time of strategies stability compared with the static pricing strategy. The fluctuation range of each party's strategies and the stability of strategy evolution is little affected by the initial cooperation rate. This also means that this pricing method can reduce the trial and error cost of setting the discharge price for the microgrid, and reduce the losses of all parties in the process of EV-MG interaction.
- (3) The simulation method used in this paper not only verifies the implementation effect of different discharge price pricing strategies in the process of EV-MG interaction, but also conducts reliable analysis in combination with specific scenarios. It provides an effective solution for studying the strategies of electric vehicle discharge electricity price in the process of EV-MG interaction.

However, this paper mainly focuses on the benefits in the process of EV-MG interaction and does not take the influence of other objective factors into consideration. In reality, there are too many uncertain factors affecting the interaction between electric vehicles and microgrids, such as charging time, weather, distance between EVs and MG and etc. Therefore, in the future, different relevant parameters can be set to conduct further research on the interaction process of EV-MG.

Acknowledgements. Research work in this paper is supported by the National Natural Science Foundation of China (Grant No. 71871160) and Shanghai Science and Technology Innovation Action Plan (No.19DZ1206800).

References

1. Xiaolin, F., Hong, W., Zhijie, W.: Research on smart transaction and collaborative scheduling strategies of micro-grid based on block-chain. *Electrical Measure. Instrumentation* pp. 1–11 (2020)
2. Sitong, W.: The new energy microgrid is sailing - the national energy administration issued the “guiding opinions on promoting the construction of new energy microgrid demonstration projects”. *Electrical Appliance Industry* **10**, 67+8 (2015)
3. Xue, H., Dongming, R., Runqing, H.: Development status and challenges of distributed renewable energy power generation in China. *China Energy* **41**(6), 32–36+47 (2019)
4. Zhang, F., Sun, C., Wei, W., et al.: Control strategies of electric charging station with V2G function based on DC Micro-Grid. In: 2015 Ieee First International Conference on Dc Microgrids, IEEE, Atlanta (2015)
5. Shan, C., Kaixuan, N., Mengyu, Z.: Two-layer coordination and optimal scheduling of microgrid for integrated charging, swapping and storage power station based on Stackelberg game. *Electric Power Automation Equipment* **40**(6), 49–59+69 (2020)
6. Rathor, S., Saxena, D.: Decentralized energy management system for LV microgrid using stochastic dynamic programming with game theory approach under stochastic environment. *IEEE Trans. Ind. Appl.* **57**(4), 3990–4000 (2021)
7. Hongbo, C., Zhicheng, L., Xun, W., et al.: Research on two-way interaction strategies of vehicle and network based on evolutionary game. *China Electric Power* **52**(07), 40–46 (2019)
8. Lin, G., Feng, X., Lu, S., et al.: Revenue optimization strategies of V2G based on evolutionary game. *J. Southeast Univ.* **36**(1), 50–55 (2020)
9. Jingdong, X., Zeyuan, B., Zhiwei, L., et al.: Supervision mechanism of transmission and distribution price adjustment for microgrid access to public distribution network services based on three-party evolutionary game. *Science Technol. Eng.* **21**(15), 6312–6321 (2021)
10. Zeng, Y., Chen, W.: The socially optimal energy storage incentives for microgrid: a real option game-theoretic approach. *Sci. Total Environ.* **710**, 36199 (2020)
11. Xianjia, W., Cuiling, G., Jinhua, Z., Ji, Q.: A review of research on stochastic evolution dynamics and its cooperation mechanism. *Syst. Sci. Mathematics* **39**(10), 1533–1552 (2019)
12. Gang, W., Yuechao, C., Yong, C., et al.: A comprehensive review of research works based on evolutionary game theory for sustainable energy development. *Energy Rep.* **8**, 114–136 (2022)
13. Vilchez, J.J.G., Jochem, P.: Simulating vehicle fleet composition: A review of system dynamics models. *Renewable and Sustainable Energy Rev.* **115**, 109367 (2019)
14. Quanlong, L., Xinchun, L., Xianfei, M.: Effectiveness research on the multi-player evolutionary game of coal-mine safety regulation in China based on system dynamics. *Safety Science* **111**, 224–233 (2019)
15. Quanlong, L., Xinchun, L.: Effective stability control of evolutionary game in China’s coal mine safety supervision supervision. *J. Beijing Inst. Technol.* **17**(04), 49–56 (2015)



An Efficient Multi-objective Evolutionary Algorithm for a Practical Dynamic Pickup and Delivery Problem

Junchuang Cai, Qingling Zhu, Qiuzhen Lin, Jianqiang Li, Jianyong Chen,
and Zhong Ming^(✉)

College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen 510860, China

caijunchuang2020@email.szu.edu.cn, {zhuqingling, qiuzhlin, lijq,
jychen, mingz}@szu.edu.cn

Abstract. Recently, practical dynamic pickup and delivery problem (DPDP) has become a challenging problem in manufacturing enterprises, due to the uncertainties of customers' requirements and production processes. This paper proposes a multi-objective evolutionary algorithm based on decomposition with four efficient local search strategies, called MOEA/D-ES, which can well solve a practical DPDP with constraints like dock, time windows, capacity and last-in-first-out loading. This method decomposes the problem under consideration into many subproblems. The experimental results on 40 real-world logistics problem instances, offered by Huawei in the competition at ICAPS 2021, validate the high efficiency and effectiveness of our proposed method.

Keywords: Dynamic pickup and delivery problem · Decomposition

1 Introduction

Dynamic Pickup and Delivery Problem (DPDP) is a fundamental problem in manufacturing enterprises, e.g., Amazon and Huawei. A lot of cargoes need to be delivered among factories during the manufacturing process. Due to the uncertainty of requirements and production processes in DPDP, most of the delivery requirements cannot be predetermined. The orders, containing information about pickup sites, delivery sites, time requirement and amount of cargoes, are generated dynamically, and then a fleet of vehicles is periodically scheduled to serve these orders. Even a small improvement of the logistics efficiency can bring significant benefits for manufacturing enterprises. However, it is very difficult to find an optimal schedule solution for DPDP due to the dynamic characteristics in these tremendous orders.

Theoretically, DPDP is a variant of traditional Vehicle Routing Problem (VRP) that is one of the famous combinatorial optimization problems. It has been proven that the VPR is one of the NP-hard problems [1]. The objective of DPDP is to dynamically dispatch each order to the most appropriate vehicle, so that the overall transportation

cost (e.g., total distances) could be minimized. Generally, the state-of-the-art methods for solving DPDP can be roughly divided into two categories. The first kind is the traditional optimization methods based on operational research [2], which decomposes the target DPDP into a series of static problems, and then each static problem is optimized using exact algorithm. The other kind is heuristic algorithms, dynamically inserting new delivery requires into scheduled routes. The heuristic algorithms include tabu search [3], cheapest insertion [4], genetic algorithm [5], waiting and buffering strategy [6].

The DPDP variant with various practical constraints is playing an increasingly important role in real-life scenarios. Unfortunately, there are few literatures paying attention to this DPDP variant. Recently, a new benchmark [7] set (HW benchmarks) is proposed by Huawei in the competition at ICAPS 2021. The HW benchmarks introduce a new practical DPDP variant with the constraints of dock, capacity, time windows and LIFO loading, which makes the problem more realistic and more challenging.

In the competition at ICAPS 2021, three teams outperformed all other 150 teams to win the prizes, whose algorithms are the state-of-the-art methods for solving the HW benchmarks. However, three teams all utilize single-objective heuristic methods to solve the practical DPDP variant and their solutions may easily get trapped in local optimum due to lacking diversity. Thus, we solved the problem using the decomposition-based multi-objective evolutionary algorithm (MOEA/D) with four efficient local search strategy. As an important research area of multi-objective evolutionary computation, MOEA/D [8] is a breakthrough of design. Its main ideas include problem decomposition, weighted aggregation of objectives, dedicated weight vectors to subproblems, and mating restriction. A detailed introduction of MOEA/D can be found in the seminal paper [8] and the surveys [9, 10]. MOEA/D has been widely used to solve many practical engineering optimization problems, such as fuzzy classifier design [11], hybrid flow shop scheduling [12], and wireless sensor network coverage design [13]. However, as MOEA/D is a generic framework, it is non-trivial to appropriately instantiate it for solving a specific problem. To design an effective MOEA/D algorithm, some application domain knowledge should be employed in designing its different components, e.g., initialization, crossover, and local search.

In summary, this paper proposes a new multi-objective evolutionary algorithm based on MOEA/D with four efficient local search strategies, called MOEA/D-ES, for solving the HW problems. A novel crossover operator and several local search strategies are adopted to simultaneously enable effective exploration and efficient exploitation. Specifically, in each re-optimization period, initial population is generated by applying the cheapest insert method on several random sequences of new orders. In MOEA/D-ES, each iteration randomly selects two parents from the population and an offspring solution is generated by crossover operator. Next, the offspring solution is then optimized by four efficient local search strategies. Besides, the ideal points and neighborhoods are updated by the offspring solution.

The remaining parts of this paper are organized as follows. Section 2 introduces the related works about DPDPs. The practical DPDP problem is introduced in Sect. 3. Section 4 introduces the proposed MOEA/D-ES. Section 4 provides the empirical results of the proposed MOEA/D-ES compared with three state-of-the-art heuristic methods. Finally, the conclusions and future works are given in the Sect. 5.

2 Preliminary

2.1 A Brief Review of DPDPs

Dynamic Pickup and Delivery Problem widely exists in many real-life applications, such as restaurant meal delivery services and door-to-door transportation services. A general overview of PDP can be found in [14, 15]. The approaches for solving the DPDP can be classified into the two following categories.

The first kind of approach for DPDP is based on traditional optimization algorithms, which solve the static problem (i.e., PDP) each time when new order requests are revealed. In [2], a planning module is firstly proposed, where the problem is decomposed into a series of static optimization problems with a subset of known delivery order over a rolling-horizon framework. The main idea of this algorithm is that it solves the static optimization problem by branch-and-price heuristics, which modifies its objective function in order to improve the overall performance in a dynamic environment. In [16], a stochastic and dynamic model for the DPDP is presented, in which demands for service arrive according to a Poisson process in time. However, the above methods present a high computation complexity, since they need to consecutively optimize each static subproblem.

The second kind of approach for DPDP is heuristic methods with local search or prediction, which are widely used in the existing studies [14]. Regarding the hybridization of heuristic algorithms with local search, a rolling-horizon was developed in [17] for the problem. The problem aims to solve the same-day pickup and delivery requests for the transport of letters and small parcels. Each of the static problems is optimized by a two-phase heuristic: the first one uses a tabu search algorithm while the second one adopts the cheapest insertion procedure. Moreover, in [6], two strategies are applied to deal with general DPDP together with a constructive-deconstructive heuristic, which shows the advantages in reducing the lost requests and the number of vehicles. Other promising methods in this kind are also showed to deal with DPDPs, such as the dial-a-flight problem about transporting passengers [18, 19], the bookings of Air taxi [20], and the meal delivery routing problem [21, 22]. However, the above algorithms cannot directly solve the HW benchmarks.

3 Problem Definition

The DPDP from the practical logistics scenarios will be described in this section. Figure 1 shows a simple example to illustrate the problem more intuitively. For this problem, our goal is to dispatch all orders to a fleet of homogeneous vehicles, aiming to obtain minimum total timeout of orders and average traveling distance of vehicles.

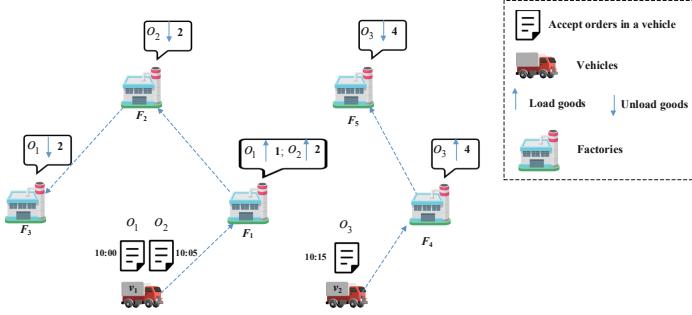


Fig. 1. A simple example of DPDp

3.1 Objective Functions

This problem includes two optimization objectives. The first objective is to minimize the total timeout of orders denoted as f_1 . If the arrival time of the order o_i to be delivered to the destination is denoted as a_d^i , then

$$f_1 = \sum_{i=1}^M \max(0, a_d^i - t_l^i), \quad (1)$$

where t_l^i is the committed completion time, and M is the total numbers of orders.

The second objective is to minimize the average traveling distance of vehicles denoted as f_2 . If the route plan of vehicle v_k is $rp_k = \{n_1^k, n_2^k, \dots, n_{l_k}^k\}$, where n_i^k stands for the i -th factory in the route of vehicle v_k , and l_k is the total numbers of nodes traveled by vehicle v_k . Then,

$$f_2 = \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^{l_k-1} d_{n_i^k, n_{i+1}^k}, \quad (2)$$

where $d_{n_i^k, n_{i+1}^k}$ is the distance from node n_i^k to node n_{i+1}^k .

3.2 Constraints

The DPDp is subject to the following constraints:

- 1) Time window constraint. Each order o_i is supposed to be served within its earliest and committed completion time. If the order is completed after the committed completion time, the penalty cost would happen; 2) The LIFO loading constraint. Each cargo must be loaded and unloaded according to the LIFO principle. This is to say, the cargoes are always placed at the rear of the vehicle when loading. Similarly, unloading at a delivery customer is allowed only if the cargoes in the current delivery are at the rear; 3) Capacity constraint. For each vehicle k , the total capacity of loaded cargoes cannot exceed the maximum loading capacity Q ; 4) Dock constraint. The number of cargo docks at each factory is limited. Each factory has a fixed number of docks. Once all the docks of the factory are occupied, the vehicles arriving later have to wait for the dock to be free in the queue of arrival order.

4 The Proposed MOEA/D-ES

4.1 Framework

When the new orders are dynamically come in the current phase, MOEA/D-ES is run to get optimized solutions. The MOEA/D [8] is a popular decomposition-based multi-objective evolutionary algorithm. It decomposes a multi-objective problem into N single-objective subproblems via a set of weight vectors $\lambda^1, \dots, \lambda^N$. A neighborhood is generated for each subproblem and different subproblems cooperate with each other during the optimization. Here, the objective function for each subproblem is:

$$f = \min \lambda_1^i \times f_1 + \lambda_2^i \times f_2, \quad (3)$$

where λ_1^i and λ_2^i are the weight of the i -th subproblem on f_1 and f_2 , respectively.

Algorithm 1 formulates the overall framework of our proposed algorithm. MOEA/D-ES first enters the initialization process (lines 1–3), including generating the T closest weight vector of each subproblem, and initializing the population and reference point. Here, the Cheapest Insert (CI) method is applied on N random sequences of new orders to generate N solutions, which formulates the initial population. Please note that the evaluation criterion is Eq. (3) when using the CI to initialize the solution of each subproblem, which ensures the diversity of the initial population. The stopping criteria (line 4) are 1) number of iterations and 2) the maximum running time. After the initialization, the algorithm evolves through iteratively performing the crossover, local search, and updating the reference point and neighborhood solutions (line 4–12). The detailed description of crossover and local search are shown in the following sections. The reference point is updated by the minimum f_1 and f_2 of offspring solution while the neighborhood solutions are updated by offspring solution thorough the Tchebycheff approach. Finally, select the best solution based on the total cost of the simulator for updating the order information in the next phase (line 13).

Algorithm 1: The General Framework of MOEA/D-ES

Input: A DPDP instance; N : the population size; m : the objective size
output: the final solution x^*

1. Generate the T closest weight vector $B(i) = \lambda^{i_1}, \dots, \lambda^{i_T}$ to each subproblem $i = 1, \dots, N$;
2. Initialize the population $P = \{x^1, \dots, x^N\}$;
3. Initialize the reference point $z = (z_1, \dots, z_m)^T$;
4. **while** Stopping criteria not satisfied **do**
5. **for** $i = 1, \dots, N$ **do**
6. Randomly select two parents p_1, p_2 from P ;
7. $x_{child} \leftarrow \text{Crossover}(p_1, p_2)$; // **Algorithm 2**
8. $x_{child} \leftarrow \text{Local_Search}(x_{child}, \lambda^i)$; // **Algorithm 3**
9. Update the reference point $z = (z_1, \dots, z_m)^T$;
10. Update of neighboring solutions;
11. **end**
12. **end while**
13. $x^* \leftarrow \text{Select_Best}(P)$; // select the best solution according to total cost at (4)
14. Return x^* ;

4.2 The Crossover Operator

In MOEAD-ES, we design a novel crossover operator for the practical DPDP, as presented in Algorithm 2. The crossover operator repeatedly selects a random route of every vehicle from each parent solution in turns and copies the selected route into the offspring solution x_{child} . Meanwhile, the duplicate nodes in x_{child} are deleted to repair x_{child} as a feasible solution. The above two steps are run repeatedly until all routes of vehicles are copied into x_{child} (lines 2–5). After that, the remaining unassigned orders (if there exist) will be inserted into x_{child} to form a complete solution. Specifically, we first save all the remaining unassigned orders in a set U (line 6). Next, for each order o_i in the U , insert the order o_i to x_{child} in the position with the minimum value of the total cost (lines 7–9). This evaluation criterion enables x_{child} to be located around the global optimal region so that the local search procedure can be easier to further exploit the global optimum. Finally, a complete solution x_{child} is formulated by the crossover operator.

Algorithm 2: The Crossover Operator

Input: parent solution p_1, p_2 , λ^i :the weight of i -th subproblem;
 K : the number of all vehicles
output: x_{child}

1. $x_{child} \leftarrow \emptyset$;
2. **for** $k = 1$ to K **do**
3. Randomly copy the route r_k from p_1, p_2 to solution x_{child} ;
4. Repair x_{child} by deleting the duplicate nodes;
5. **end**
6. $U \leftarrow$ all the remaining unassigned orders;
7. **for** each order $o_i \in U$ **do**
8. Insert o_i to x_{child} with the minimum value of TC at (4);
9. **end**
10. Return x_{child} ;

4.3 The Local Search Procedure

As shown in lines 1–24 of Algorithm 3, the local search procedure is core component of MOEA/D-ES which consists of four local search strategies: couple-exchange, block-exchange, block-relocate, and couple-relocate. These operators can be referred to [23, 24]. Specifically, the couple-exchange swap two couples (e.g., couple $[x^+, x^-]$ and couple $[w^+, w^-]$) to create a new route (Fig. 2(b)); the block-exchange swap two blocks (e.g., block B_x and block B_w) to create a new route (Fig. 2(c)); the couple-relocate relocate a couple (e.g., couple $[z^+, z^-]$) to another feasible position to create a new route (Fig. 3(b)); the block-relocate relocate a couple (e.g., block B_x) to another feasible position to create a new route (Fig. 3(c)).

Algorithm 3: The Local Search Procedure

Input: offspring solution x_{child} , λ^i :the weight of the i -th solution;
output: a solution x_{ls}

1. **Repeat**
2. $x_1 \leftarrow \text{Couple-Exchange } (x_{child})$
3. **if** $f(x_1) < f(x_{child})$ **then**
4. $x_{child} \leftarrow x_1$;
5. **continue**;
6. **end if**
7. $x_1 \leftarrow \text{Block-Exchange } (x_{child})$
8. **if** $f(x_1) < f(x_{child})$ **then**
9. $x_{child} \leftarrow x_1$;
10. **continue**;
11. **end if**
12. $x_1 \leftarrow \text{Block-Relocate } (x_{child})$
13. **if** $f(x_1) < f(x_{child})$ **then**
14. $x_{child} \leftarrow x_1$;
15. **continue**;
16. **end if**
17. $x_1 \leftarrow \text{Relocate-Couple } (x_{child})$
18. **if** $f(x_1) < f(x_{child})$ **then**
19. $x_{child} \leftarrow x_1$;
20. **continue**;
21. **else**
22. **break**;
23. **end if**
24. **until** $used_time > MAX_TIME$;
25. $x_{ls} \leftarrow x_{child}$;
26. **return** x_{ls} ;

Initially, the first neighborhood couple-exchange strategy is continuously applied to perform local search around x_{child} to find an improved solution x_1 with a smaller value f at Eq. (3), and update it to x_{child} (line 2–6). If no improved solution is generated, the second neighborhood block-exchange strategy is applied to perform local search around x_{child} , also trying to find an improved solution of x_{child} . x_{child} is updated if there is an improved solution, and the local search procedure jumps back to the first neighborhood couple-exchange strategy to continuously search for an improved solution of x_{child} (line 7–11). Otherwise, run the third neighborhood block-relocate strategy to find an improved solution of x_{child} . If a better solution is found, block-relocate strategy updates x_{child} and jumps back to the first neighborhood strategy (line 12–16). At last, run the fourth neighborhood block-relocate strategy to perform similar operations as before (line 17–20). It is considered that the solution x_{child} has reached the optimum status in the four neighborhoods when an improved solution x_1 cannot be found after the fourth neighborhood strategy. The evaluation function f at Eq. (3), containing different weight λ^i in different iterations, enables more adequate coverage on the space of feasible solutions around x_{child} . Consequently, the local search procedure can effectively avoid falling into the local optimal region. Meanwhile, the time limit for the algorithm to solve

a batch of orders is 10 min and the orders will be dynamically coming in each of 10 min. This is a hard constraint due to the practical requirement of solving the DPDP.

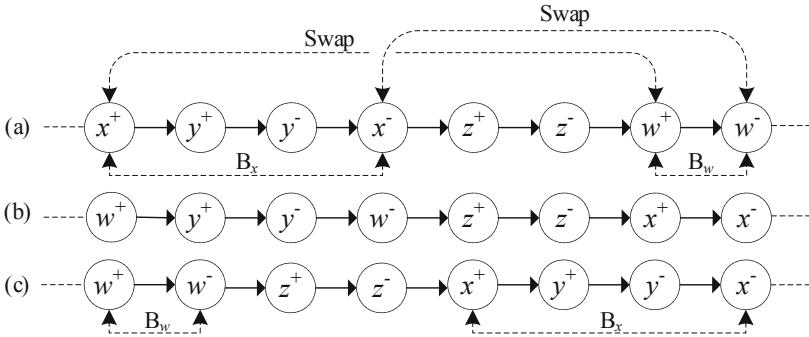


Fig. 2. (a) The initial route. (b) The new route created by couple-exchange. (c) The new route created by block-exchange.

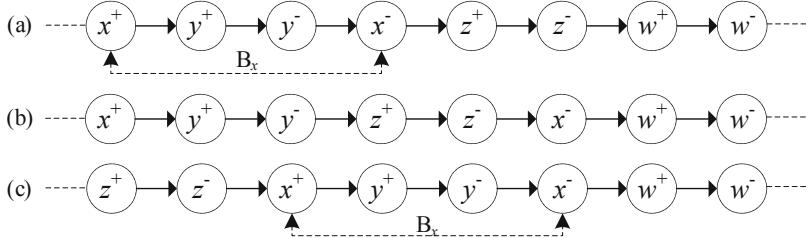


Fig. 3. (a) The initial route. (b) The new route created by couple-relocate. (c) The new route created by block-relocate.

5 Experimental Study

5.1 Benchmark Problems and Performance Metrics

The DPDP variant with practical constraints is very common in industrial scene. The HW benchmarks [7], proposed by Huawei in the competition at ICAPS 2021, introduce a new practical DPDP variant with practical enterprise-scale constraints of dock, time windows, capacity and LIFO loading, which makes the problem more realistic and more difficult. The datasets from Huawei's real system contain 30 days of historical data. As illustrated in Table 1, the HW benchmarks contain three homogeneous vehicle quantity scales (5, 20, and 50) and five order quantity scales (50, 100, 300, 500, and 1000).

Table 1. The number of vehicles and orders of different HW instances

Instances	Vehicle numbers	Order numbers
HW1 - HW8	5	50
HW9 - HW16	5	100
HW17 - HW24	20	300
HW25 - HW32	20	500
HW33 - HW40	50	1000

In each HW instance, there are multiple factories that can dynamically generate delivery orders within a day. In the HW benchmarks, the goal is to dispatch all orders to a fleet of vehicles, aiming to obtain minimum total timeout of orders and average traveling distance of vehicles. Therefore, the performance metric is the total cost:

$$TC = \alpha \times f_1 + f_2, \quad (4)$$

where α is a large positive constant, showing that the enterprises are more concerned about timely fulfillment of orders in practical scenario to improve customers' satisfaction.

5.2 The Compared Algorithms and Experimental Settings

To verify the effectiveness of the proposed MOEA/D-ES, three competitive state-of-the-art heuristic algorithms for solving the HW benchmarks are included for comparison, including gold-winning, silver-winning and bronze-winning algorithms in ICAPS 2021 [25]. A brief introduction of each compared algorithm is given below:

- 1) The gold-winning algorithm: The algorithm utilized variable neighborhood search method for finding out the route plans of vehicles. Generally, it continuously swaps the pickup nodes group and delivery nodes group among route plans of different vehicles and also exchange the orders inside a route plan for better result. The pickup (delivery) nodes group are adjacent pickup (delivery) nodes with same address. Moreover, the delay strategy is applied by this algorithm to get the better solution.
- 2) The silver-winning algorithm: This algorithm adopts a strategy that transforms the DPDP problem into a knapsack problem and packs nodes for each route taking all practical constraints into consideration at the same time. Specifically, it is developed from the perspective of threshold check, i.e., allocating the orders by checking whether they reach the threshold of delivery time and vehicle capacity. If so, those orders would be allocated.
- 3) The bronze-winning algorithm: The algorithm constructs the initial solution for the DPDP by the CI algorithm, in which the pickup node and delivery node are consecutive in the route for every insertion operation. Moreover, the solution is improved by local search through a ruin-reconstruct strategy.

The running time of all the compared algorithms and MOEA/D-ES to process the orders in each interval period is limited to 10 min. MOEA/D-ES and all compared

algorithms run each instance of the HW benchmarks ten times independently, reserving the best total cost at Eq. (4) of each algorithm.

The source codes of the gold-winning algorithm, the silver-winning algorithm, and the bronze-winning algorithm can be found from the official competition website [26]. These algorithms have been tuned to the best by their authors for solving the same set of problems adopted in this paper, which can ensure a fair comparison. The common settings of our experiments are 1) penalty value for one hour tardiness: $\alpha = 10000/3600$, 2) maximum number of docks in each factory is set to 6, 3) maximum loading capacity Q is adopted as 16, and 4) maximum running time is set to 10 min. These settings are consistent with the settings of the DPDP competition at ICAPS 2021.

Please note that some unique parameter settings for each compared algorithm are set as suggested in their corresponding references. In addition, some unique parameter settings of MOEA/D-ES for all test problems are introduced as below:

- 1) Population size (N): The population size is set to 6, as the problem is of time consuming.
- 2) Neighborhood size (T): The neighborhood size is adopted as 2.
- 3) *Stopping criteria* includes: I) the maximum number of iterations ($50 \times N$) and II) the maximum running time (600 s)

5.3 The Comparison Experiments on the HW Benchmarks

Table 2 presents the best and average performance of MOEA/D-ES and the three compared algorithms on the HW benchmarks (HW1-HW40). The column headed “Gap” in Table 2 presents the gap between f_{\min} and f_{best} , which is equal to $(f_{\min} - f_{best})/f_{best}$. Here, f_{\min} and f_{best} are respectively the solution with minimum TC obtained by an algorithm on an instance across 10 runs and the best solution of the instance found by all algorithms across 10 runs.

It is obvious that the proposed MOEA/D-ES showed better overall performance than the other three competitors. Specifically, MOEA/D-ES can obtain the best results on 31 out of the 40 instances, while the gold-winning algorithm, the sliver-winning algorithm, and the bronze-winning algorithm only show the best results on 8, 1 and 0 out of 40 instances. From the one-by-one comparisons in the row ‘best/all’ of Table 2, MOEA/D-ES performed better than the gold-winning algorithm, the sliver-winning algorithm, and the bronze-winning algorithm in 32, 39 and 40 out of 40 instances, respectively, and it was only respectively worse in 8, 1 and 0 out of 40 instances. It is worth mentioning that the gap of the 9 instances, on which MOEA/D-ES performs worse than the other three competitors, is almost small, indicating the robustness of MOEA/D-ES. Besides, from the ‘Avg’ perspective, the average performance of MOEA/D-ES is 6.21E+3 while that of the other three competitors are 1.14E+4, 1.48E+5 and 3.89E+4, respectively, which also shows the outstanding performance of MOEA/D-ES.

Table 2. Comparison between MOEA/D-ES, Gold-winning, Silver-winning, and Bronze-winning algorithms on the HW benchmarks in ten runs.

Instance	MOEA/D-ES		Gold		Silver		Bronze	
	MTC	Gap	MTC	Gap	MTC	Gap	MTC	Gap
HW1	1.17E + 2	0.0000	1.34E + 2	0.1449	2.30E + 3	18.584	1.30E + 2	0.1069
HW2	8.88E + 1	0.0000	9.56E + 1	0.0766	3.05E + 4	342.46	9.14E + 1	0.0293
HW3	9.41E + 1	0.0000	9.68E + 1	0.0283	3.58E + 4	379.44	9.65E + 1	0.0255
HW4	9.45E + 1	0.0000	9.46E + 1	0.0008	5.49E + 3	57.095	1.04E + 2	0.1005
HW5	3.31E + 3	0.0004	3.31E + 3	0.0000	1.69E + 4	4.1124	5.45E + 3	0.6487
HW6	1.05E + 2	0.0000	1.05E + 2	0.0029	4.76E + 3	44.445	1.18E + 2	0.1266
HW7	4.32E + 3	0.0000	4.39E + 3	0.0160	1.28E + 4	1.9626	7.36E + 3	0.7035
HW8	6.39E + 1	0.0000	6.88E + 1	0.0780	7.97E + 2	11.480	7.69E + 2	11.042
HW9	1.65E + 2	0.0868	1.52E + 2	0.0000	1.81E + 5	1188.0	8.48E + 3	54.708
HW10	7.46E + 4	0.0000	1.94E + 5	1.6020	1.77E + 6	22.711	1.87E + 5	1.5051
HW11	1.61E + 2	0.0000	1.98E + 2	0.2291	1.81E + 5	1123.9	3.04E + 3	17.893
HW12	8.61E + 3	0.0000	5.29E + 4	5.1503	5.67E + 5	64.861	8.42E + 4	8.7805
HW13	1.73E + 2	0.0000	7.18E + 3	40.587	2.20E + 5	1273.7	2.77E + 2	0.6051
HW14	1.50E + 2	0.0000	9.39E + 3	61.739	2.37E + 5	1582.1	7.82E + 3	51.237
HW15	1.92E + 4	0.4144	1.35E + 4	0.0000	7.93E + 5	57.539	1.49E + 5	9.9993
HW16	1.04E + 4	0.0000	1.68E + 4	0.6072	8.54E + 5	80.901	5.78E + 4	4.5432
HW17	7.34E + 1	0.0000	8.17E + 1	0.1139	9.65E + 1	0.3153	4.11E + 2	4.6017
HW18	7.97E + 1	0.0000	8.22E + 1	0.0318	1.12E + 2	0.4060	8.56E + 3	106.45
HW19	1.06E + 2	0.0000	1.09E + 2	0.0219	4.78E + 2	3.4999	4.15E + 3	38.068
HW20	3.30E + 3	0.0001	3.30E + 3	0.0000	3.98E + 3	0.2074	1.62E + 4	3.9146
HW21	9.74E + 1	0.0000	1.13E + 2	0.1641	2.34E + 3	23.037	1.85E + 4	189.03
HW22	1.64E + 3	0.0000	1.65E + 3	0.0054	3.32E + 3	1.0289	2.11E + 4	11.894
HW23	1.06E + 2	0.0059	1.05E + 2	0.0000	6.87E + 3	64.198	8.09E + 2	6.6777
HW24	8.53E + 1	0.0000	9.44E + 1	0.1069	1.29E + 3	14.123	2.00E + 3	22.446
HW25	5.98E + 3	0.0000	1.03E + 4	0.7144	9.24E + 4	14.444	2.15E + 4	2.5937
HW26	4.48E + 3	0.0000	8.95E + 3	0.9951	2.75E + 5	60.319	5.40E + 4	11.041
HW27	1.19E + 2	0.0000	1.34E + 2	0.1240	1.84E + 4	153.26	1.88E + 4	156.61
HW28	7.12E + 3	0.0031	7.10E + 3	0.0000	9.44E + 3	0.3293	1.49E + 4	1.0982
HW29	5.92E + 3	0.0000	6.40E + 3	0.0804	1.63E + 5	26.522	4.24E + 4	6.1593
HW30	1.10E + 2	0.0000	1.19E + 2	0.0809	7.40E + 4	671.54	2.11E + 4	190.76
HW31	1.45E + 4	0.0000	2.26E + 4	0.5609	1.47E + 5	9.1543	2.33E + 4	0.6095
HW32	5.69E + 3	0.0000	7.67E + 3	0.3484	6.04E + 4	9.6176	1.96E + 4	2.4455
HW33	1.39E + 3	0.0000	1.59E + 3	0.1445	7.03E + 3	4.0660	9.50E + 4	67.459
HW34	6.32E + 3	0.0000	1.05E + 4	0.6596	1.05E + 4	0.6603	5.44E + 4	7.6021
HW35	8.12E + 1	0.0000	3.44E + 3	41.35	1.58E + 4	193.47	1.04E + 5	1279.0

(continued)

Table 2. (*continued*)

Instance	MOEA/D-ES		Gold		Silver		Bronze	
	<i>MTC</i>	Gap	<i>MTC</i>	Gap	<i>MTC</i>	Gap	<i>MTC</i>	Gap
HW36	1.34E + 4	0.0000	1.75E + 4	0.3038	2.32E + 4	0.7264	1.13E + 5	7.4088
HW37	1.15E + 4	1.2293	1.12E + 4	1.1845	5.14E + 3	0.0000	9.57E + 4	17.618
HW38	1.42E + 4	0.0000	1.50E + 4	0.0617	3.01E + 4	1.1243	9.42E + 4	5.6482
HW39	1.74E + 4	0.1675	1.49E + 4	0.0000	2.23E + 4	0.4947	8.52E + 4	4.7107
HW40	1.31E + 4	0.2881	1.02E + 4	0.0000	2.43E + 4	1.3819	1.17E + 5	10.468
<i>best/all</i>	31/40		8/40		1/40		0/40	
Avg	6.21E + 3	0.0000	1.14E + 4	0.8346	1.48E + 5	22.779	3.89E + 4	5.2681

The reason why MOEA/D-ES outperforms the other three compared algorithms significantly is that MOEA/D-ES can well balance the exploration and exploitation compared to the other three competitors. Since the other three competitors use only one solution throughout the whole optimization period, the solutions can easily fall into local optimal region. On the contrary, MOEA/D-ES first initializes N solutions of population with different weights, which ensures the diversity of the population. Besides, the evaluation criterion TC at Eq. (4) in crossover procedure enables offspring solution to locate around the global optimal region. Moreover, the weight vectors in overall objective function Eq. (3) applied to different offspring solutions are different, making the solutions jump out of local optimality easily. What's more, the population in MOEA/D-ES can quickly converge to the optimal solution region, as different subproblems cooperate with each other during the procedures like updating reference point and neighboring solutions.

In summary, MOEA/D-ES is of high efficiency when dealing with the HW benchmarks. The remarkable advantages of MOEA/D-ES are due to the fact that, VNSME has a strong ability in exploring feasible search space since it can gain high-diversity solutions through independent exploitation and collaboration of N subproblems. As a result, VNSME can usually identify more promising regions than the compared algorithms, which has higher probability for finding superior solutions.

6 Conclusions and Future Work

This paper proposes a multi-objective evolutionary algorithm based on MOEA/D with four efficient local search strategies, called MOEA/D-ES, for solving a practical DPDP with constraints like dock, time windows, capacity and last-in-first-out loading. This method decomposes the target DPDP under consideration into many subproblems. As a result, the experimental results on 40 real-world logistics problem instances, offered by Huawei in the competition at ICAPS 2021, validate the high efficiency and effectiveness of our proposed method.

In the future, we plan to add the function of the delayed delivery and the forecast strategy about orders to MOEA/D-ES, trying to further improve its performance in this new practical DPD. Besides, apply the deep reinforcement learning methods to solve the DPD variant can be another further research direction.

References

1. Cordeau, J.-F., Laporte, G., Ropke, S.: Recent Models and Algorithms for One-to-One Pickup and Delivery Problems. pp. 327–357. Springer US https://doi.org/10.1007/978-0-387-77778-8_15
2. Savelsbergh, M., Sol, M.: Drive: dynamic routing of independent vehicles. *Oper. Res.* **46**(4), 474–490 (1998)
3. Gendreau, M., Guertin, F., Potvin, J.-Y., Séguin, R.: Neighborhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries. *Trans. Res. Part C: Emerging Technol.* **14**(3), 157–174 (2006)
4. Mitrović-Minić, S., Laporte, G.: Waiting strategies for the dynamic pickup and delivery problem with time windows. *Trans. Res. Part B: Methodological* **38**(7), 635–655 (2004)
5. D. Sáez, C. E. Cortés, A. Núñez, O. Research: Hybrid adaptive predictive control for the multi-vehicle dynamic pick-up and delivery problem based on genetic algorithms and fuzzy clustering. *Comput. Oper. Res.* **35**(11), 3412–3438 (2008)
6. Pureza, V., Laporte, G.: Waiting and buffering strategies for the dynamic pickup and delivery problem with time windows. *INFOR: Information Systems Operational Res.* **46**(3), 165–175 (2008)
7. Hao, J., Lu, J., Li, X., Tong, X., Xiang, X., Yuan, M., Zhuo, H.H.: Introduction to the dynamic pickup and delivery problem benchmark--ICAPS 2021 competition. arXiv preprint [arXiv: 2202.01256](https://arxiv.org/abs/2202.01256) (2022)
8. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **11**(6), 712–731 (2007)
9. Trivedi, A., Srinivasan, D., Sanyal, K., Ghosh, A.: A survey of multiobjective evolutionary algorithms based on decomposition. *IEEE Trans. Evol. Comput.* **21**(3), 440–462 (2016)
10. Xu, Q., Xu, Z., Ma, T.: A survey of multiobjective evolutionary algorithms based on decomposition: variants, challenges and future directions. *IEEE Access* **8**, 41588–41614 (2020)
11. Zhang, Q., Li, B., Zhang, F.: A MOEA/D approach to exploit the crucial structure of convolution kernels. In: 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI), pp. 643–648
12. Jiang, S.-L., Zhang, L.: Energy-oriented scheduling for hybrid flow shop with limited buffers through efficient multi-objective optimization. *IEEE Access* **7**, 34477–34487 (2019)
13. Xu, Y., Ding, O., Qu, R., Li, K.: Hybrid multi-objective evolutionary algorithms based on decomposition for wireless sensor network coverage optimization. *Appl. Soft Comput.* **68**, 268–282 (2018)
14. Berbeglia, G., Cordeau, J.-F., Laporte, G.: Dynamic pickup and delivery problems. *Eur. J. Oper. Res.* **202**(1), 8–15 (2010)
15. Psaraftis, H.N., Wen, M., Kontovas, C.A.: Dynamic vehicle routing problems: three decades and counting. *Networks* **67**(1), 3–31 (2016)
16. Swihart, M.R., Papastavrou, J.D.: A stochastic and dynamic model for the single-vehicle pick-up and delivery problem. *Eur. J. Oper. Res.* **114**(3), 447–464 (1999)
17. Mitrović-Minić, S., Krishnamurti, R., Laporte, G.: Double-horizon based heuristics for the dynamic pickup and delivery problem with time windows. *Trans. Res. Part B: Methodological* **38**(8), 669–685 (2004)

18. Schilde, M., Doerner, K.F., Hartl, R.F.: Metaheuristics for the dynamic stochastic dial-a-ride problem with expected return transports. *Comput. Oper. Res.* **38**(12), 1719–1730 (2011)
19. Schilde, M., Doerner, K.F., Hartl, R.F.: Integrating stochastic time-dependent travel speed in solution methods for the dynamic dial-a-ride problem. *Eur. J. Oper. Res.* **238**(1), 18–30 (2014)
20. Fagerholt, K., Foss, B., Horgen, O.: A decision support model for establishing an air taxi service: a case study. *J. Operational Res. Soc.* **60**(9), 1173–1182 (2009)
21. Reyes, D., Erera, A., Savelsbergh, M., Sahasrabudhe, S., O’Neil, R.: The meal delivery routing problem. *Optimization Online* (2018)
22. Ulmer, M.W., Thomas, B.W., Campbell, A.M., Woyak, N.: The restaurant meal delivery problem: dynamic pickup and delivery with deadlines and random ready times. *Transp. Sci.* **55**(1), 75–100 (2021)
23. Cassani, L., Righini, G.: Heuristic algorithms for the TSP with rear-loading. In: 35th Annual Conference of the Italian Operational Research Society (AIRO XXXV), Lecce (2004)
24. Carrabs, F., Cordeau, J.-F., Laporte, G.: Variable neighborhood search for the pickup and delivery traveling salesman problem with LIFO loading. *INFORMS J. Comput.* **19**(4), 618–632 (2007)
25. <https://competition.huaweicloud.com/information/1000041411/circumstance>
26. <https://competition.huaweicloud.com/information/1000041411/Winning>



An Efficient Evaluation Mechanism for Evolutionary Reinforcement Learning

Xiaoqiang Wu, Qingling Zhu, Qiuzhen Lin^(✉), Jianqiang Li, Jianyong Chen,
and Zhong Ming

College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen 510860, China

2110276189@email.szu.edu.cn, {zhuqingling,qiuzhlin,lijq,jychen,
mingz}@szu.edu.cn

Abstract. In recent years, many algorithms use Evolutionary Algorithms (EAs) to help Reinforcement Learning (RL) jump out of local optima. Evolutionary Reinforcement Learning (ERL) is a popular algorithm in this field. However, ERL evaluate the population in each loop of the algorithm, which is inefficient because of the uncertainties of the population's experience. In this paper, we propose a novel evaluation mechanism, which only evaluates the population when the RL agent has difficulty in studying further. This mechanism can improve the efficiency of the hybrid algorithms in most cases, and even in the worst scenario, it only reduces the performance marginally. We embed this mechanism into ERL, denoted as E-ERL, and compare it with original ERL and other state-of-the-art RL algorithms. Results on six continuous control problems validate the efficiency of our method.

Keywords: Evolutionary algorithm · Reinforcement learning · Evolutionary reinforcement learning · Evaluation mechanism

1 Introduction

Reinforcement learning (RL) has reached or surpassed human level in a number of different problem domains. Deep Q-network (DQN) achieved a level comparable to that of a professional human games tester in Atari [1]. In addition, Alpha Go [2] and OpenAI Five [3] defeated the world champion in Go and Dota2 respectively. An important reason why reinforcement learning can achieve so much is the combination with deep neural networks, which is generally referred to as Deep Reinforcement Learning (DRL). Although DRL has been successfully applied to various problem areas, it is still not common to use DRL to solve real-world problems. One important reason is the lack of effective exploration in DRL and premature convergence to local optima [4].

Evolutionary algorithms (EAs) are a class of black-box optimization techniques, which maintain a population of individuals to search for the optimal solution. In [5], Genetic Algorithms (GAs), a kind of EAs, have been proven to be a competitive alternative in solving many RL problems. Evolution Strategies (ES) [6], another branch of EAs,

also show competitive results on most Atari games [7]. And both of algorithms are very suitable for parallel implementation, because only a small amount of information needs to be exchanged in computing. EAs are generally considered to have strong exploration ability due to the diversity of populations and they are tolerant of sparse rewards because the cumulative rewards of a complete episode are selected as individuals' fitness. Also for this reason, EAs only leverage the information of the sum of rewards in one episode and ignore the specific information in every step, they generally suffer from lower sample efficiency than RL methods.

To combine the advantages of these two methods and make up for their shortcomings, many hybrid algorithms which consist of RL and EA algorithm are proposed in recent years. Khadka and Tumer [4] integrate these two approaches in their Evolutionary Reinforcement Learning (ERL) framework. ERL maintain a population of agents and an extra RL agent, and the RL agent is copied into the population periodically. In the ERL framework, the population and the RL agent are evaluated in every loop of the algorithm and then the RL agent learns the experience from the population and the RL agent. One of the main purposes of evaluating the population is using the experience from the population to help RL agent learn and escape from local optima. However, the experience of the population is not always better than RL agent's. When population's experience is not as useful as RL agent's, it's inefficient to evaluate the whole population. Besides, in some real world domains, such as self-driving vehicles [7] and aerial delivery [8], interacting with the environment is expensive and unsafe, unnecessary evaluations are more costly than other areas. To improve the efficiency of hybrid algorithms, we design a novel evaluation mechanism, which only evaluates the population when the RL agent cannot make a further progress. This evaluation mechanism can be applied to ERL and other similar hybrid algorithms. The main contributions of this paper are introduced as follows:

- 1) An effective evaluation mechanism is proposed to enhance the efficiency of ERL and other similar hybrid algorithms.
- 2) We embed this evaluation mechanism into ERL, called E-ERL. The experimental results validate the effectiveness of the proposed evaluation mechanism when compared to the original ERL and other state-of-the-art RL algorithms on solving the continuous control tasks from Mujoco environments [9].

The remaining parts of this paper are organized as follows. Section 2 introduces the related works which about the combination of RL and EA algorithms. Section 3 introduces the details of the proposed evaluation mechanism and the implementation in ERL. Section 4 provides the empirical results of the proposed E-ERL and other algorithms. Finally, the conclusions and future works are given in the Sect. 5.

2 Preliminary

2.1 Related Works of the Combination of EA and RL

Merging Evolutionary Algorithms and Deep Reinforcement Learning to solve complex RL problems is an emerging trend. These hybrid algorithms benefit from the advantages of both kinds of methods and can generally achieve better performances than separate EA or RL.

A popular hybrid algorithm that combines policy gradient and neuroevolution is ERL. There is a population of agents and a separate RL agent in the ERL framework. The individuals in the population and the RL agent optimize their policy networks by neuroevolution and gradient method respectively. ERL evaluates all individuals in population and the RL agent in every loop of the algorithm and the RL agent is inserted into the population periodically.

Proximal Distilled Evolutionary Reinforcement Learning (PDERL) [11] extends ERL with different crossover and mutation operators. PDERL uses Q-filtered distillation crossover and Proximal mutations to avoid the catastrophic forgetting and damaging effects caused by n-points crossover and Gaussian noise. CEM-RL [12] proposes a different combination scheme using the simple cross-entropy method (CEM) [13] to replace ERL's neuroevolution part. CERL [14] extends ERL to distributed training with multiple different actor and computational resources are dynamically allocated. MERL [15] extends ERL to solve multiple agent reinforcement learning problems. MERL uses EA and MADDPG [16] to optimize the team reward and agent-specific reward respectively. C2HRL [17] uses a heterogeneous agent pool to leverage advantages of different agents. Specifically, C2HRL's agent pool includes TD3 [18], SAC [19] and EA agents.

3 The Proposed Evaluation Mechanism and E-ERL

3.1 The Proposed Evaluation Mechanism

ERL evaluates all individuals in population and RL agent in every loop and stores $N + 1$ agents' experience in the replay buffer, where N is the size of the population. Then the RL agent samples a random minibatch from the replay buffer and uses it to update its parameters using gradient descent. However, the experience from the individuals of population is not always better than the experience from the RL agent, especially when the crossover and mutation operators of EA algorithm are totally random. If the experience of the population is worse than RL agents', evaluating every individual in population and save their experience for learning is inefficient. Therefore, we design an evaluation mechanism that the population will be evaluated only when the RL agent fall into the local optima. Only in this situation, we evaluate the whole population and expect the population's experience can help RL agent escape from the local optima. In practical problems, it's hard to confirm if the RL agent converges prematurely, so we use a simple method to determine whether the RL agent needs the help from the population: we only evaluate the population when the RL agent cannot make progress for K consecutive times.

Algorithm 1: The pseudo-code of the proposed E-ERL

```

1 Initialize population  $pop_{\pi}$  and RL agent, replay buffer R,  $score_m = 0$ ,  $k = 0$ 
2 while True do
3    $score = \text{Evaluate}(\pi_{rl}, R, \xi)$ 
4   if  $score < score_m$  then
5     if  $k \geq K$  then
6       for actor  $\pi \in pop_{\pi}$  do:
7          $fitness, R = \text{Evaluate}(\pi, R, \xi)$ 
8       Generate the next generation of population
9     else
10     $k = k + 1$ 
11  end if
12 else
13    $score_m = score$ 
14    $k = 0$ 
15   replace the worst individual in population with  $\pi_{rl}$ 
16 end if
17   RL agent updates parameters by gradient method
18 end while

```

Algorithm 2: Function Evaluate

```

1 procedure Evaluate( $\pi, R, \xi$ )
2    $fitness = 0$ 
3   for  $i = 1 : \xi$  do
4     Reset environment and get initial state  $s_0$ 
5     while env is not done do
6       Select action  $a_t = \pi(s_t | \theta^{\pi})$ 
7       Execute action  $a_t$  and get reward  $r_t$  and new state  $s_{t+1}$ 
8       Append transition  $(s_t, a_t, r_t, s_{t+1})$  to R
9        $fitness = fitness + r_t$  and  $s = s_{t+1}$ 
10      end while
11    end for
12    Return  $fitness / \xi, R$ 
13 end procedure

```

3.2 The Proposed E-ERL

We embed above introduced evaluation mechanism into the framework of ERL and we dub it as E-ERL. The pseudo-code of the proposed E-ERL is given in Algorithm 1 and Algorithm 2. For the genetic operators used in the EA part of the algorithm, we follow the implementation published by the author of ERL.

We use variable k to control whether to evaluate the population. At the start of the algorithm, k will be set to 0. If the RL agent makes no further progress in this loop, k will be set to $k + 1$, otherwise k will be set to 0. Once k is not smaller than K , the population will be evaluated and its experience will be stored in the replay buffer. In ERL, the RL agent is inserted into the EA population periodically, this synchronization mode is not applicable to our method because of the change of evaluation mechanism.

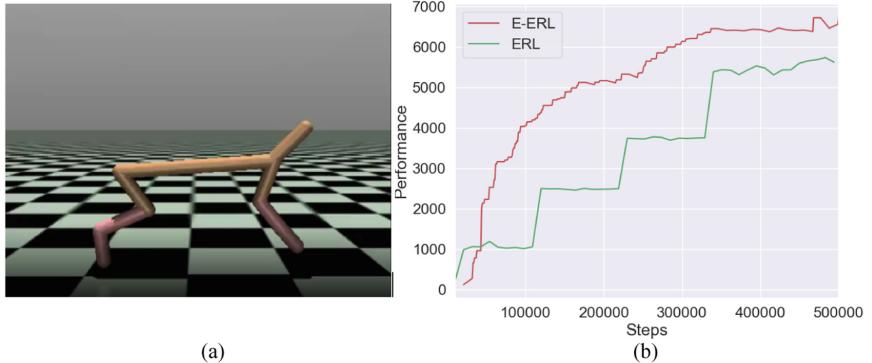


Fig. 1. (a) The HalfCheetah environment; (b) Learning curves of ERL and E-ERL in the first 0.5 million steps

Therefore, we delete this periodical synchronization and copy the RL agent into the population when the RL agent breaks the best record of the scores. Furthermore, because the sum of rewards in one episode is used to measure the performance of the RL agent in E-ERL, we also evaluate the RL agent for ξ times and use the average. ξ varies in different environment and we use the same value as ERL in all environments. In order to further measure accurately the performance of the RL agent, we do not use the exploration noise in RL algorithm. It might weaken the exploration of RL algorithm, but the diverse experience of the population can reduce greatly the influence of this.

In extreme circumstances, if RL agent can make progress all the time, the population will have no effect in E-ERL. It's reasonable when the RL agent can perform well by itself. On the contrary, if the advances of the RL agent rely on the population totally, the behaviors of E-ERL are similar as ERL. In other words, even in the worst scenario for the proposed E-ERL, its performance still approaches ERL.

3.3 Discussions on E-ERL

Figure 1(a) shows the HalfCheetah environment in Mujoco. The agent needs to control a biped robot to move forward in this environment. The performance of agents in HalfCheetah is relatively stable, therefore, we utilize this environment to illustrate the performance difference between ERL and E-ERL. In ERL, the population size and the synchronization period used in the HalfCheetah are set to 10. The length of one episode in HalfCheetah is 1000 steps, thus the length of one loop in ERL algorithm is 11000 steps because of the evaluation of the population and the RL agent. Periodically, the RL agent is copied into the population every 110000 steps. As can be observed in Fig. 1(b), the performance of the best individual in the population increases significantly only when the RL agent's insertion happens and changes slightly at other stages. This phenomenon indicates the new individuals produced in the population do not make much progress and the increase of performance is attributed to the learning of the RL agent. The experiences of the population might help the RL agent in some cases, but not all the time. Therefore, we design a different evaluation mechanism and improve the efficiency of the hybrid

algorithm. And the results in Fig. 1(b) show the effect of our method, the RL agent can learn faster after omitting the redundant evaluations.

4 Experimental Studies

In this section, several experiments are conducted to investigate the performance of the proposed E-ERL in solving robot locomotion problems. First, the effectiveness of the proposed evaluation mechanism is empirically studied. Then the experimental comparisons between the proposed E-ERL and several popular algorithms (i.e., ERL [4], PPO [20], DDPG [21]) on solving continuous control problems are conducted.

4.1 Benchmark Problems and Performance Metrics

Mujoco environments are used widely in many field [5, 22-24] and wrapped by the OpenAI gym [25] to provide interface for the researchers and engineers. We compare the performance of different algorithms on 6 robot locomotion problems powered by Mujoco. These 6 test problems are HalfCheetah, Swimmer, Hopper, Reacher, Ant, and Walker-2d. The goal of these problems is to apply a torque on the joints to make the robot finish some particular actions.

The cumulative reward in one episode is used as the performance of agents and the average in multiple episodes is calculated in a number of unstable environments. For ERL and E-ERL, the performance of the best individual in the population is regarded as the performance of the algorithm. To compare fairly between RL algorithms and population-based algorithms, the steps of every individual in population will be cumulated.

4.2 The Compared Algorithms and Experimental Settings

To verify the effectiveness of the proposed evaluation mechanism, we implement it in ERL and compare our method to three competitive algorithms for solving continuous control problems, including ERL [4], DDPG [21] and PPO [20]. A brief introduction of each compared algorithm is given below:

- 1) ERL [4]: ERL is a popular hybrid algorithm that combines policy gradient and neuroevolution and it's the first algorithm as the combination of the two approaches to achieve practical benefits in robot locomotion tasks [11].
- 2) DDPG [18]: Deep Deterministic Policy Gradient is a widely used policy gradient method in domains with continuous action space and it's the RL algorithm adopted in ERL.
- 3) PPO [20]: PPO is a state-of-the-art on-policy RL algorithm and can be used for solving large scale real-world problems [26].

All unique parameter settings in each compared algorithm are set as suggested in their corresponding references. We use the implementation in Open AI Baselines [27] for PPO and DDPG, and ERL's implementation published by its author. In addition, some common parameter settings for ERL and E-ERL are the same. The added hyperparameter K in E-ERL is 10 in our experiments.

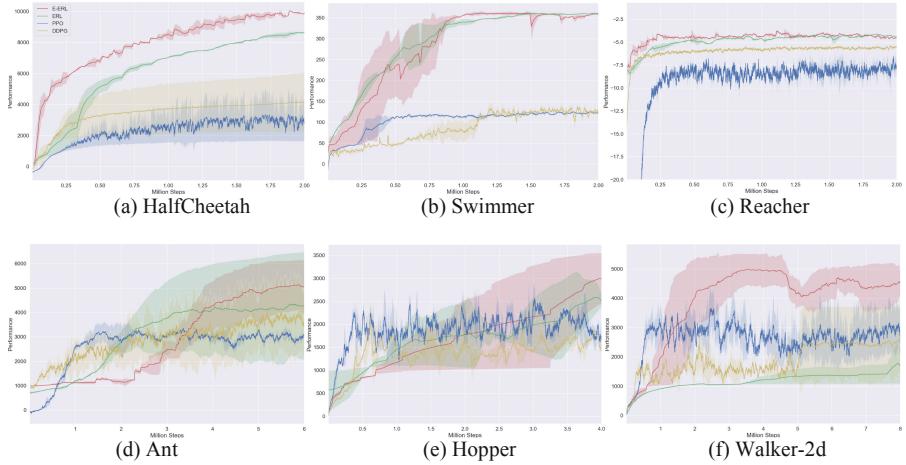


Fig. 2. Learning curves on six Mujoco continuous control benchmarks

4.3 Comparisons Between E-ERL and Other Algorithms

Figure 2 provides the comparative results of E-ERL, ERL, DDPG and PPO. As can be observed from Fig. 2, E-ERL outperforms DDPG and PPO in all environments. Compared to ERL, E-ERL obtains improvements in almost all environments except Ant and shows competitive results in Ant. E-ERL outperforms ERL at early stages in HalfCheetah, Reacher and Walker2d, because the evaluation mechanism reduces useless evaluations and accelerates the learning process. Besides, E-ERL achieves significantly better final results in HalfCheetah and Walker2d. For ERL, the increase of agents' performance depends on the iteration of population in Swimmer and Hopper. However, E-ERL evaluates the population less frequently and the number of population's iterations decreases. Therefore, the results of E-ERL in the early period are not as well as ERL when the environment is Swimmer or Hopper. But the ultimate performance of E-ERL will catch up with and even exceed ERL's in these two problems. Specifically, the Swimmer benchmark is difficult for RL algorithms to learn but is relatively easier for EA [4], so PPO and DDPG perform poorly in this environment. Apart from the improvement in the cumulative reward, the computation burden in E-ERL is smaller than ERL. Because ERL executes more crossover and mutation operations when the number of timesteps is same.

5 Conclusions and Future Work

In this paper, an efficient evaluation mechanism is presented for improving the efficiency of the hybrid algorithms combining EA and RL, which only evaluates all individuals in population when the RL agent falls into the local optima. This mechanism has been embedded into the framework of ERL, namely E-ERL. At the beginning of the proposed E-ERL, a variable k and the record of the RL agent's best score $score_m$ are initialized. Then, if the RL agent cannot make progress for K consecutive times, the whole population

will be evaluated and the experience of the population will be stored into the replay buffer for subsequent learning. Otherwise, we only evaluate the and update the RL agent, nothing will happen to population. In this way, many useless evaluations are omitted and the quality of experience in the replay buffer is improved. Hence, the efficiency in ERL is strengthened by using the proposed evaluation mechanism, which is crucial for solving real-world problems. When compared to ERL, DDPG and PPO, E-ERL showed an obvious superiority over other competitors when solving the continuous control problems.

Regarding our future work, 1) the application in other hybrid algorithms will be studied in our future work. 2) the performance of E-ERL on solving other continuous control problems and the combination with other RL algorithms [28–31] or EA algorithms [32–35] will be further studied. Moreover, 3) the selection of valuable individuals that deserved to be evaluated and the exploration of more effective evaluation mechanisms will also be considered in our future work.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grants 61876110, 61836005, and 61672358, the Joint Funds of the National Natural Science Foundation of China under Key Program Grant U1713212, and Shenzhen Technology Plan under Grant JCYJ20190808164211203.

References

1. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)
2. Silver, D., et al.: Mastering the game of go without human knowledge. *Nature* **550**, 354–359 (2017)
3. Berner, C., et al.: Dota 2 with large scale deep reinforcement learning. arXiv preprint [arXiv:1912.06680](https://arxiv.org/abs/1912.06680) (2019)
4. Khadka, S., Tumer, K.: Evolution-guided policy gradient in reinforcement learning. In: *Advances in Neural Information Processing Systems*. pp. 1188–1200 (2018)
5. Such, F.P., Madhavan, V., Conti, E., Lehman, J., Stanley, K.O., Clune, J.: Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. arXiv preprint [arXiv:1712.06567](https://arxiv.org/abs/1712.06567) (2017)
6. Rechenberg, I., Eigen, M.: *Evolutionsstrategie: Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*. Frommann-Holzboog Stuttgart (1973)
7. Salimans, T., Ho, J., Chen, X., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. ArXiv e-prints (2017)
8. Burnett, K., Qian, J., Du, X., Liu, L., Yoon, D.J., et al.: Zeus: A system description of the twotime winner of the collegiate SAE autodrive competition. *J. Field Robotics* **38**, 139–166 (2021)
9. Boutilier, J.J., Brooks, S.C., Janmohamed, A., Byers, A., Buick, J.E., et al.: Optimizing a drone network to deliver automated external defibrillators. *Circulation* **135**, 2454–2465 (2017)
10. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pp. 5026–5033. IEEE (2012)
11. Bodnar, C., Day, B., Lió, P.: Proximal distilled evolutionary reinforcement learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI'20)*, pp. 3283–3290. AAAI Press (2020)

12. Pourchot, A., Sigaud, O.: CEM-RL: Combining evolutionary and gradient-based methods for policy search. In: International Conference on Learning Representations (2019)
13. De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. *Annals of Operations Res.* **134**(1), 19–67 (2005)
14. Khadka, S., et al.: Collaborative evolutionary reinforcement learning. In: International Conference on Machine Learning (2019)
15. Majumdar, S., Khadka, S., Miret, S., McAlleer, S., Tumer, K.: Evolutionary reinforcement learning for sample-efficient multiagent coordination. In: International Conference on Machine Learning, pp. 6651–6660 (2020)
16. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O.P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative competitive environments. In: Advances in Neural Information Processing Systems, pp. 6379–6390 (2017)
17. Zheng, H., Jiang, J., Wei, P., Long, G., Zhang, C.: Competitive and cooperative heterogeneous deep reinforcement learning. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS’20, pp. 1656–1664, Richland, SC (2020)
18. Fujimoto, S., van Hoof, H., Meger, D.: Addressing function approximation error in actor-critic methods. In: International Conference on Machine Learning (2018)
19. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning (2018)
20. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint [arXiv:1707.06347](https://arxiv.org/abs/1707.06347) (2017)
21. Lillicrap, T.P., et al.: Continuous control with deep reinforcement learning. arXiv preprint [arXiv:1509.02971](https://arxiv.org/abs/1509.02971) (2015)
22. Duan, Y., Chen, X., Houthooft, R., Schulman, J., Abbeel, P.: Benchmarking deep reinforcement learning for continuous control. In: International Conference on Machine Learning, pp. 1329–1338 (2016)
23. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. arXiv preprint [arXiv:1709.06560](https://arxiv.org/abs/1709.06560) (2017)
24. Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust region policy optimization. In: International Conference on Machine Learning, pp. 1889–1897 (2015)
25. Brockman, G., et al.: Openai gym. arXiv preprint [arXiv:1606.01540](https://arxiv.org/abs/1606.01540) (2016)
26. Ye, D., et al.: Mastering complex control in MOBA games with deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6672–6679 (2020)
27. Dhariwal, P., et al.: Openai Baselines. <https://github.com/openai/baselines> (2017)
28. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937 (2016)
29. Cobbe, K.W., Hilton, J., Klimov, O., Schulman, J.: Phasic policy gradient. In: International Conference on Machine Learning, PMLR, pp. 2020–2027 (2021)
30. Hasselt, V.H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2094–2100 (2016)
31. Hessel, M., et al.: Rainbow: Combining improvements in deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
32. Conti, E., Madhavan, V., Such, F.P., Lehman, J., Stanley, K.O., Clune, J.: Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. arXiv preprint [arXiv:1712.06560](https://arxiv.org/abs/1712.06560) (2017)
33. Cully, A., Clune, J., Tarapore, D., Mouret, J.-B.: Robots that can adapt like animals. *Nature* **521**(7553), 503 (2015)

34. Lehman, J., Stanley, K.O.: Exploiting open-endedness to solve problems through the search for novelty. In: ALIFE, pp. 329–336 (2008)
35. Pugh, J.K., Soros, L.B., Stanley, K.O.: Quality diversity: a new frontier for evolutionary computation. *Frontiers in Robotics and AI* **3**, 40 (2016)



A Mixed-Factor Evolutionary Algorithm for Multi-objective Knapsack Problem

Yanlian Du^{1,2}, Zejing Feng³, and Yijun Shen²(✉)

¹ School of Information and Communication Engineering, Hainan University,
Haikou 570228, China

² State Key Laboratory of Marine Resources Utilization in South China Sea, Hainan University,
Haikou 570228, China
sheny2000@hainanu.edu.cn

³ College of Applied Science and Technology, Hainan University, Danzhou 571737, China

Abstract. Nondominated-sorting plays an important role in multi-objective evolutionary algorithm in recent decades. However, it fails to work well when the target multi-objective problem has a complex Pareto front, brusque nondominated-sorting virtually steers by the conflicting nature of objectives, which leads to irrationality. In this paper, a novel mixed-factor evolutionary algorithm is proposed. A normalization procedure, i.e. mixed-factor, is introduced in the objective space, which links all the objectives for all the solutions of the problem to ease the conflicting nature. In the process of nondominated-sorting, the mixed factors of individual substitute the raw objectives. In order to ensure that the population are thoroughly steered through the normalized objective space, hybrid ageing operator and static hypermutation with first constructive mutation are used to guide the searching agents converge towards the true Pareto front. The algorithm proposed is operated on multi-objective knapsack problem. The effectiveness of MFEA is compared with five state-of-the-art algorithms, i.e., NSGA-II, NSGA-III, MOEA/D, SPEA2 and GrEA, in terms of five performance metrics. Simulation results demonstrate that MFEA achieves better performance.

Keywords: Multi-objective evolutionary algorithm · Mixed-factor evolutionary algorithm · Multi-objective Knapsack Problem · Normalization procedure

1 Introduction

In the real world, multi-objective optimization problem (MOP) is ubiquitous in decision making, it is common to face a problem with several conflicting objectives. Over the past decades, a number of multi-objective evolutionary algorithms (MOEAs) have been suggested for MOPs, mainly for the reason that, MOEAs can find a set of trade-off Pareto-optimal solutions in a single run. In 1985, vector evaluated genetic algorithms (VEGA) was initiated by D. Schaffer [1], who is considered to be the first one to design a MOEA. Since then, many successful MOEAs have been developed, e.g., PAES [2], SPEA2 [3], IBEA [4], MOEA/D [5, 6], NSGA-II [6], NSGA-III [7], GrEA [8], etc. Today, MOEAs are successfully applied to resolve engineering and real-life problems,

such as feature selection [9], virtual machine scheduling [10], the location problem of wireless sensor nodes [11] and so on.

In general, the goal of MOPs is to find a variety of nondominated solutions when the execution of the algorithm is terminated, notably, to examine the trajectory of Pareto front. In most dominance-based MOEAs, nondominated-sorting is used from the very start and will go through the whole generations. However, early in the iteration, a population of strictly nondominated solutions is not enough to represent the direction of Pareto front. This is because those are more likely solutions that are randomly generated or just represent part of the Pareto front. Strictly nondominated-sorting may force the search to a bias Pareto front. Thus, a mechanism of trajectory is needed to be developed to steer the population towards the right Pareto front. The challenge of multi-objective optimization (MOO) is how to construct a comprehensive evaluation mechanism which links all the objectives of all the solutions. In terms of MOEAs, the key problem lies in two aspects: 1) constructing a selection mechanism that assigns for each individual a probability of selection proportionally to its multiple objectives; 2) the selection mechanism is constructed based on the multiple objectives, with the aim to overcome or ease the conflicting nature of the objectives. Nonetheless, Pareto dominance-based fitness evaluation fails to generate such a positive guidance mechanism in two aspects: 1) while solutions in the current population are mostly nondominated (actually they stay away from the true Pareto front), and 2) when solutions are next to be nondominated, the performance of Pareto dominance-based selection is unstable.

In this work, a novel normalization procedure, i.e., mixed-factor, is introduced, which is used to operate on the objective space before comparing the candidate solutions. In order to overcome the slow convergence and local optima of MOEA, and ensure that the population are appropriately steered through the normalized objective space, hybrid ageing operator and static hypermutation with first constructive mutation (FCM) [12] are used to guide the searching agents converge towards the true Pareto front.

As one of the classical combinatorial optimization problems, Multi-objective Knapsack Problem (MOKP) is applied in a wide range of fields, such as information technology, industrial engineering, economic management, etc. Herein, the proposed algorithm is assessed by MOKP. The rest of this work is organized as follows: in the next section, some basic concepts of MOEA and MOKP are reviewed. In Sect. 3, the suggested MFEA is presented with detailed definitions and procedure. In Sect. 4, the experimental results obtained from the suggested algorithm and other five state-of-the-art MOEAs are discussed. Finally, a conclusion is drawn and a plan of the future work is proposed in Sect. 5.

2 Background

2.1 Multi-objective Evolutionary Algorithm

Different from single-objective optimization problems, which aims at achieving an optimal solution that hits the only criterion, generally, MOPs involve several conflicting objectives. The optimality of MOPs is usually defined by Pareto dominance, where a solution is defined as optimal if it cannot be dominated by all other solutions. E.g., in a multi-objective minimization problem, the goal is to find all solutions $x \in \Omega$ such that

the vector-valued objective function $F(x) = (f_1(x), f_2(x), \dots, f_k(x))$ is Pareto optimal. As follows (1) and (2), $x \in \Omega$ is a decision vector (feasible solution) in the decision space Ω , k is the number of objective functions, $\prod_{i=1}^k \mathbb{D}_i$ is the feasible objective space. In practice, the decision space Ω must satisfy certain conditions (e.g., equality constraints and/or inequality constraints) according to specific problems.

$$(MOP) \begin{cases} \min F(x) = (f_1(x), f_2(x), \dots, f_m(x)) \\ s.t. \quad x \in \Omega \end{cases} \quad (1)$$

$$f_i : \Omega \rightarrow \mathbb{D}_i, i = 1, 2, \dots, k \quad (2)$$

Pareto Optimality

In a multi-objective minimization problem, a feasible solution $x \in \Omega$ is considered to dominate another feasible solution x' , denoted as $x \prec x'$, if and only if, $\forall i \in \{1, 2, \dots, k\}, f_i(x) \leq f_i(x')$ and $F(x) \neq F(x')$. A feasible solution $x^* \in \Omega$ is defined as a Pareto optimal, if and only if, $\nexists y \in \Omega$ such that $y \prec x^*$ [13]. All Pareto optimal solutions make up a set called Pareto optimal set (PS) (3). The evaluation of PS is called the Pareto front (PF) (4).

$$PS = \{x^* \in \Omega | \nexists y \in \Omega, y \prec x^*\} \quad (3)$$

$$PF = \{F(x^*) | x^* \in PS\} \quad (4)$$

From the view point of MOEA, Pareto-optimal solutions represent phenotype characteristics, while the corresponding objectives act as genotype information whose components cannot be all optimized synchronously because of the conflicting nature.

2.2 Multi-objective Knapsack Problem

As one of the classical combinatorial optimization problems, MOKP [14] has been studied extensively. It is proven to be NP – hard [15], in addition, the size of the Pareto optimal solutions set may grow with the number of objectives, and items in the knapsack dramatically [16]. Despite its concise model, MOKP can be widely employed to resolve many engineering and real-life issues, e.g., budget and resource allocation [17] etc. Hence, its resolution has attracted the attention of scholars for theoretical and practical research. Mathematically, MOKP can be stated as follows (5) [14]: given n items, p arguments $w_j^i, i \in \{1, 2, \dots, p\}, j \in \{1, 2, \dots, n\}$ (price, height, weight, volume, etc.), and k profits, $c_j^i, i \in \{1, \dots, k\}, j \in \{1, \dots, n\}$, how to select items from the n ones, aiming at maximizing the k profits, without exceeding the p knapsack capacities W_i ?

$$\begin{cases} \max f_i(x) = \sum_{j=1}^n c_j^i x_j, i \in \{1, 2, \dots, k\} \\ s.t. \sum_{j=1}^n w_j^i x_j \leq W_i, i \in \{1, 2, \dots, p\} \\ x_j \in \{0, 1\}, \forall j \in \{1, 2, \dots, n\} \end{cases} \quad (5)$$

3 Mixed-Factor Evolutionary Algorithm

In Pareto dominance, objectives are conflicting as in nature, i.e., an improvement in one objective is at the cost of deterioration of another. Generally, the main goal of MOPs is to find a variety of nondominated solutions. However, when solve MOPs based on Pareto dominance, brusque nondominated-sorting virtually steer by the conflicting nature of objectives (genotype information), which leads to irrationality. Hence, more attention should be on the preprocess of original objective space which aim at easing the conflicting nature before nondominated-sorting. In other words, a mechanism of comprehensive assessment is needed to be developed to normalize the original objective. In most dominance-based MOEAs, nondominated-sorting is used from the very start and through the whole generations. However, in the early generations, a population of strictly nondominated solutions is not enough to represent the direction of Pareto front. This is because those solutions are probably generated randomly or just part of the Pareto front, which are far from the true Pareto front. The difficulty of MOO is how to construct a comprehensive evaluation mechanism which links all the objectives of all the solutions. In terms of MOEAs, the key problems to be resolved are two aspects: 1) constructing a selection mechanism that assigns for each individual a probability of selection proportionally to its multiple objectives; 2) the selection mechanism is constructed based on the multiple objectives, with the aim at overcoming or easing the conflicting nature of the objectives. In this section, a novel normalization procedure, i.e., mixed-factor is proposed to preprocess the objective space before dominated-sorting, the objective space is replaced by the normalized mixed-factor in nondominated-sorting procedure. In order to ensure that the population are thoroughly steered through the normalized objective space, hybrid ageing operator and static hypermutation with first constructive mutation (FCM) are used to guide the searching agents converge towards the true Pareto front.

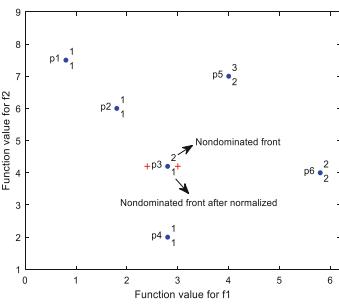


Fig. 1. Pareto front

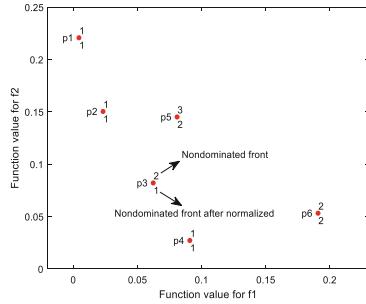


Fig. 2. Pareto front after normalization

3.1 Definitions

Before comparing the candidate solutions, it is crucial to preprocess the objective space of the population. As in Fig. 1 of a minimization problem, from the notion of Pareto

dominance, individuals $\{p_1, p_2, p_4\}$ belong to the first nondominated front (NF), while $\{p_3, p_6\}$ belongs to the second front, $\{p_5\}$ belongs to the third front. However, as Fig. 1 shows, p_3 is capricious, if impose a tiny decrement on the first objective of p_3 , e.g., move to left '+' in Fig. 1, p_3 switches to the first front, otherwise, stay or move a bit to the right, p_3 belongs to the second front. This turbulence weakens the ability of nondominated-sorting. On the contrary, it implies the fact that, the left '+' in Fig. 1 belongs to the first front though it is not discovered in the current iteration, which conveys that: 1) the left '+' in Fig. 1 belongs to the first front, however, it is not discovered in the current iteration; 2) hence p_3 should be added to the first front as it locates in a strategic position, which greatly record the trajectory of Pareto front; 3) the relative objective values are more significant than the absolute ones, moreover, the relative objective value in the population (outer factor E_{ij} in this work) and each relative objective value in each individual (inner factor I_{ij}) should be integrated into account to put forward a comprehensive assessment of each objective of each individual. From this point of view, in this section, a novel normalization procedure, i.e., mixed-factor is proposed to preprocess the objective space before comparing the candidate solutions, which serves as a probability of selection proportionally to the performance of each objective.

Given an objective space Y (6), i.e., k objectives of p individuals in the population. Here suppose that $y_{ij} \geq 0, \forall i, j$, otherwise, replace y_{ij} by $\exp(y_{ij})$ to make the objective space non negative.

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pk} \end{pmatrix} \quad (6)$$

Definition 1 (Internal Factor)- The internal factor $I_{ij} = y_{ij}/y_i$. of individual p_i is the internal factor of the i th individual that make itself be chosen as descendant for the j th objective, as follows (7).

$$I = (I_{ij})_{p \times k} = \begin{pmatrix} \frac{y_{11}}{y_1} & \frac{y_{21}}{y_1} & \cdots & \frac{y_{1k}}{y_1} \\ \frac{y_{21}}{y_2} & \frac{y_{22}}{y_2} & \cdots & \frac{y_{2k}}{y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{y_{p1}}{y_p} & \frac{y_{p2}}{y_p} & \cdots & \frac{y_{pk}}{y_p} \end{pmatrix} \quad (7)$$

where $y_i = \sum_{j=1}^k y_{ij}$, $I_i = \sum_{j=1}^k I_{ij}=1, \forall i \in \{1, 2, \dots, p\}$. Conceptually, I_{ij} resembles self-competence, as each individual aware of its own skills, $I_i = 1$ conveys the idea that every individual is a complete synthesis, pursuing all-round development.

Definition 2 (External Factor)- The external factor $E_{ij} = y_{ij}/y_j$ of individual p_i is the external part of the i th individual that make itself be chosen as descendant for the j th

objective in the whole population, as follows, (8).

$$E = (E_{ij})_{p \times k} = \begin{pmatrix} \frac{y_{11}}{y_1} & \frac{y_{21}}{y_2} & \dots & \frac{y_{1k}}{y_k} \\ \frac{y_{21}}{y_1} & \frac{y_{22}}{y_2} & \dots & \frac{y_{2k}}{y_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{y_{p1}}{y_1} & \frac{y_{p2}}{y_2} & \dots & \frac{y_{pk}}{y_k} \end{pmatrix} \quad (8)$$

Here $y_j = \sum_{i=1}^p y_{ij}$, $E_j = \sum_{i=1}^k E_{ij} = 1$, $\forall j \in \{1, \dots, p\}$. Conceptually E_{ij} is similar to competitive social skill, which individuals tend to compare with each other in the environment. $E_j = 1$ reflects the fact that every individual is competitive in every objective just for its own contribution to the whole.

Definition 3 (Mixed Factor)- The mixed factor $M_{ij} = I_{ij} \oplus E_{ij}$ is a comprehensive factor of the i th individual that make itself be chosen as descendant for the j th objective in the whole population.

$$M = I \oplus E \quad (9)$$

Here “ \oplus ” is an operator which integrates the inner and outer factor to comprehensively assess the performance of individual. In order to evolve the appropriate “ \oplus ” operator, two ways are employed in this work, as (10) and (11). In (11), $q = 1 - p$, $0 \leq p \leq 1$. p reflects the weight of internal factor, while q reflects the one of external.

$$I_{ij} \oplus E_{ij} = pI_{ij} + qE_{ij} \quad (10)$$

$$I_{ij} \oplus E_{ij} = I_{ij} \times E_{ij} \quad (11)$$

In the experiments, p changes linearly from 0 to 1 by a tolerance of 0.1. On equal conditions, compared with (11), the experimental results from (10) are far worse than (11). Herein, in the remainder of this paper, (11) is used. It can be seen from Fig. 1, 2 that, firstly p_3 belongs to the second front, but after normalized, p_3 belongs to the first front, which greatly verifies its strategic location, and at the same time, records the trajectory of PF. With mixed factor, the conflicting nature of objectives is greatly eased.

3.2 Framework of MFEA

In order to overcome the slow convergence and local optima of MOEA, and to ensure that the population are thoroughly steered through the normalized objective space, herein, static hypermutation coupled with FCM is embedded in the Offspring creation procedure with a small probability rh , hybrid ageing operator is used to guide the searching agents towards the true Pareto front without getting stuck in local optima.

The main loop of MFEA is described in **Algorithm 1**. First, an initial parent population P of size N is randomly generated, the candidate offspring population P' of size N is produced using the Binary Tournament Selection (BTS). Then, Offspring creation procedure equipped with Static hypermutation (**Algorithm 2**) is employed to generate the

official offspring population, since as (Dogan Corus, et al., 2019) proved that, hypermutations with FCM can induce substantial speed-ups when local optima is to be overcome. In Static hypermutation with FCM (**Algorithm 3**), if there is no improvement achieved in the first flip, the operator will perform at most nh flips until any fitness improvement is achieved. At the end of the variation stage, the individuals newly created reset $age = 0$ only if their fitness is better than that of their parents', if not, they inherit their parents' age. Right after variation stage, the age of the whole population (including parents and offspring) is increased by one. In addition, the ageing operator eliminates old individuals by the way that individuals die out with a probability r_{die} once they reach an age of τ . It was shown that the hybrid ageing helps to escape local optima [18], therefore, we employ this operator in MFEA in this paper and give its pseudo-code in **Algorithm 4**. Finally, the objectives of the parent population and offspring population are calculated, the raw objective space is normalized before nondominated-sorting procedure is executed. In MFEA, the normalized mixed-factors M_{ij} substitute the raw objectives in the nondominated-sorting.

Algorithm 1 Pseudo Code of MFEA

```

Generate  $N$  individuals to form initial population  $P_0$ 
Set  $t = 0$ 
While (stopping conditions are not satisfied) do
     $P'_t$  = Binary Tournament Selection ( $P_t$ )
     $C_t$  = Offspring creation ( $P'_t$ )
     $R_t = C_t \cup P_t$ 
    Hybrid ageing ( $R_t$ )
    Calculate the objectives of  $R_t$ 
    Execute the normalization procedure ( $R_t$ )
     $P_{t+1}$  = Nondominated-sorting ( $R_t$ )
    Selection ( $P_{t+1}$ )
     $t = t + 1$ 
end while

```

Algorithm 2 Offspring creation (P'_t)

```

Crossover and mutation ( $P'_t$ )
Set  $i = 1$ 
for each  $p_i \in P'_t$  do
    Generate a random number  $rand$  between 0 and 1
    if  $rand < rh$ 
        Static hypermutation( $p_i$ ) with FCM
    end if
     $i = i + 1$ 
end for

```

Algorithm 3 Static hypermutation (p_i)

```

if FCM is not used then
    create  $p'_i$  by flipping  $np$  distinct bits selected uniformly at random
else
    create  $p'_i$  by flipping at most  $np$  distinct bits selected uniformly at random one after another until a constructive mutation happens
end if
if  $p'_i \prec p_i$  then
     $p_i = p'_i$ 
     $p_i^{age} = 0$ 
else
     $p_i = p'_i$ 
end if
```

Algorithm 4 Hybrid ageing (P_{t+1})

```

for all  $p_i \in P_{t+1}$  do
     $p_i^{age} = p_i^{age} + 1$ 
    if  $p_i^{age} > \tau$  then
        Remove  $p_i$  from  $P_{t+1}$  with probability  $r_{die}$ 
    end if
end for
```

Algorithm 5 Selection (P_{t+1})

```

if  $|P_{t+1}| > N$  then
    remove  $(|P_{t+1}| - N)$  individuals with the lowest fitness breaking ties uniformly at random
end if
if  $|P_{t+1}| < N$  then
    Generate  $(N - |P_{t+1}|)$  individuals uniformly at random
end if
```

3.3 Offspring Creation

For producing offspring, firstly a pool of parent candidates is generated via binary tournament selection, crossover and mutation are executed immediately after. The selection is operated solely based on PF and crowding distance (CD) [6], in the normalized objective space of mixed-factor. In the Static hypermutation, each individual is selected to undergo hypermutation randomly with a probability of rh . If no improvement is found in the first flip, the operation continues no more than np times until any fitness improvement is achieved, just as FCM says, stop at first constructive mutation. The mechanism behind it is that, in order to enhance adaptability, high mutation rates arise in the immune system. This is one of the typically important characteristics observed in the world, survival of the fittest [12]. Detail steps are presented in **Algorithm 2**.

4 Experiments and Results

4.1 Experiment Methodology

A test has been performed at the MFEA on benchmark instances of MOKP chosen from the instance libraries: Zitzler and al. [19], i.e., nine instances with the number of items 250, 500, and 750, with two, three and four objectives. Uncorrelated profits c_j^i and weights w_j^i were integers chosen randomly in the interval [10, 100]. The knapsack capacities W_i were set to half of the total weight regarding the corresponding knapsack.

The experiments are carried out on a personal computer equipped with Intel® Core™ i5-6300HQ CPU, 2.30 GHz, Memory: 2.00 GB, Windows 10. The performance of MFEA with five MOPs with different concepts and/or different search strategies are compared, i.e., SPEA2 [3], MOEA/D [6, 8], NSGA-III [7], and GrEA [8]. The parameters settings used for each algorithm are chosen according to the papers from which they originated. All algorithms are initiated with 100, 105 and 120 individuals for 2, 3, 4 objective MOKP, all are executed for 2000 generations. The results are mean of 30 independent runs.

Table 1. Number of rankings of the compared algorithms

Algorithm	Rank					
	1st	2nd	3rd	4th	5 th	6th
MFEA	23	8	1	1	9	3
NSGAI	0	3	23	10	9	0
NSGAI	0	4	3	3	14	21
MOEAD	14	22	0	0	0	9
SPEA2	7	5	11	22	0	0
GrEA	1	3	7	9	13	12

4.2 Constraints Handling

Considering that, the effect of mix-factor is tested, the performance of MOEA equipped with hybrid ageing and hypermutation with FCM is under examination, the correction procedure used in this work is nothing special. To be specific, the illegal agents flip the decision variables of value one to zero randomly, until all constraints are satisfied. Apparently, this random iterative correction procedure prevents from imposing genes (solution characteristics), or any influence on the operators of MFEA. Therefore, the performance of MFEA is thoroughly examined.

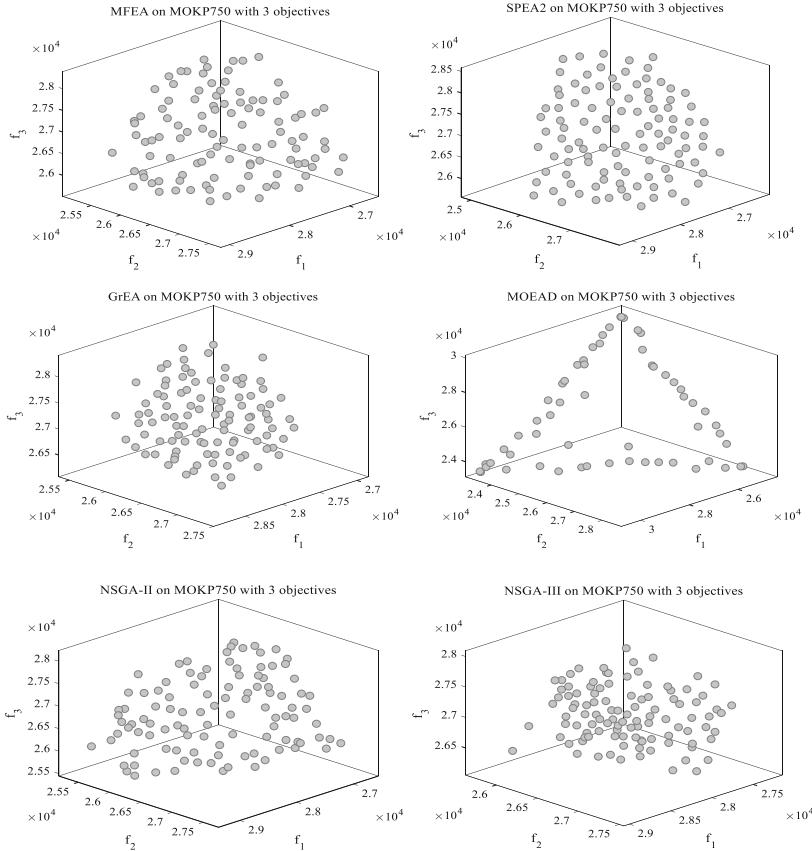


Fig. 3. Plots of non-dominated fronts obtained by 6 algorithms in the 3-objective MOKP750.

4.3 Experimental Results and Comments

In order to evaluate and compare the quality of solutions evolved by MFEA and other algorithms, five performance metrics: Inverted Generational Distance (IGD) [20], Hyper-volume (HV) [21], Spacing metric (SM) [22], Generational Distance (GD) [20], and Convergence metric (CM) [21] are used.

Table 1 demonstrates the number of each ranks of the algorithms in each metric, it is seen that MFEA is the best. Apparently outperforms NSGA-II, NSGA-III, SPEA2 and GrEA, MOEAD seems to achieve comparable performance with MFEA. However, we can see from Table 2 that, MFEA achieves the best IGD and HV in 2-objective MOKP than other algorithms, MOEAD achieves the best GD and CM. In Table 3, we can see that MFEA outperforms others in IGD, GD and CM, while MOEAD achieves the best HV, SPEA2 obtains the best SM. In Table 4, MFEA achieves the best IGD, CM and GD values among all 4-objective instances, while MOEAD obtains the best HV and SPEA2 achieves the best SM. Overall, MFEA achieve 23 times the ‘first’ in all metrics of all instances, even more than the sum of other five algorithms do. The MFEA demonstrates

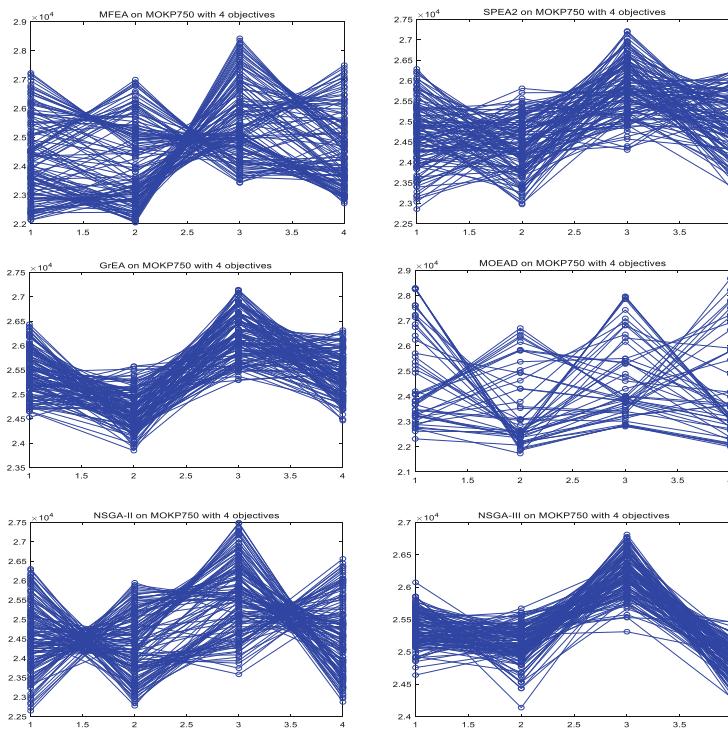


Fig. 4. Parallel coordinates of non-dominated fronts obtained by 6 algorithms in 4-objectives MOKP750.

Table 2. Comparisons of results of MFEA and other algorithms on 2-objective MOKP

Metric	Algorithm	Instance		
		2.250	2.500	2.750
		Mean	Mean	Mean
IGD	MFEA	6.51E+03	1.31E+04	2.02E+04
	NSGAII	6.90E+03	1.43E+04	2.05E+04
	NSGAIII	6.93E+03	1.43E+04	2.05E+04
	MOEAD	6.51E+03	1.33E+04	1.88E+04
	SPEA2	6.90E+03	1.42E+04	2.03E+04
	GrEA	6.92E+03	1.43E+04	2.04E+04
HV	MFEA	3.85E+07	1.47E+08	2.68E+08
	NSGAII	3.24E+07	1.25E+08	2.63E+08

(continued)

Table 2. (*continued*)

Metric	Algorithm	Instance		
		2.250	2.500	2.750
		Mean	Mean	Mean
CM	NSGAI ^{III}	3.17E+07	1.24E+08	2.54E+08
	MOEAD	3.73E+07	1.47E+08	3.24E+08
	SPEA2	3.26E+07	1.30E+08	2.69E+08
	GrEA	3.26E+07	1.26E+08	2.62E+08
	MFEA	3.42E+00	6.20E+00	8.65E+00
	NSGAI ^{II}	7.51E+00	1.23E+01	1.12E+01
GD	NSGAI ^{III}	8.55E+00	1.29E+01	1.36E+01
	MOEAD	3.74E+00	4.57E+00	3.38E+00
	SPEA2	7.45E+00	9.75E+00	9.20E+00
	GrEA	7.77E+00	1.23E+01	1.06E+01
	MFEA	6.56E+03	1.40E+04	2.02E+04
	NSGAI ^{II}	6.90E+03	1.43E+04	2.05E+04
SM	NSGAI ^{III}	6.93E+03	1.43E+04	2.05E+04
	MOEAD	6.51E+03	1.33E+04	1.88E+04
	SPEA2	6.90E+03	1.42E+04	2.03E+04
	GrEA	6.92E+03	1.43E+04	2.04E+04
	MFEA	1.28E+01	3.14E+01	4.01E+01
	NSGAI ^{II}	1.71E+01	2.97E+01	5.17E+01

to have a better IGD with a significant difference. MOEAD seems to be competitive, achieves better performance in Hypervolume than MFEA do, but the performance of IGD is far less than MFEA.

Table 3. Comparisons of results of MFEA and other algorithms on 3-objective MOKP

Metric	Algorithm	Instance		
		3.250	3.500	2.750
		Mean	Mean	Mean
IGD	MFEA	4.66E+03	1.01E+04	1.54E+04
	NSGAII	6.13E+03	1.26E+04	1.94E+04
	NSGAIII	6.57E+03	1.32E+04	2.02E+04
	MOEAD	5.56E+03	1.16E+04	1.73E+04
	SPEA2	6.25E+03	1.28E+04	1.97E+04
	GrEA	6.59E+03	1.33E+04	2.03E+04
HV	MFEA	1.13E+11	7.39E+11	2.17E+12
	NSGAII	1.14E+11	8.01E+11	2.49E+12
	NSGAIII	1.03E+11	7.24E+11	2.32E+12
	MOEAD	1.67E+11	1.11E+12	3.65E+12
	SPEA2	1.18E+11	8.12E+11	2.57E+12
	GrEA	1.07E+11	7.60E+11	2.45E+12
CM	MFEA	2.05E+00	3.02E+00	3.67E+00
	NSGAII	4.63E+00	6.08E+00	8.19E+00
	NSGAIII	7.21E+00	9.92E+00	1.32E+01
	MOEAD	2.50E+00	3.51E+00	3.43E+00
	SPEA2	4.74E+00	6.60E+00	8.40E+00
	GrEA	7.09E+00	9.38E+00	1.15E+01
GD	MFEA	4.66E+03	1.01E+04	1.54E+04
	NSGAII	6.13E+03	1.26E+04	1.94E+04
	NSGAIII	6.57E+03	1.32E+04	2.02E+04
	MOEAD	5.56E+03	1.16E+04	1.73E+04
	SPEA2	6.25E+03	1.28E+04	1.97E+04
	GrEA	6.59E+03	1.33E+04	2.03E+04
SM	MFEA	1.17E+02	1.80E+02	2.35E+02
	NSGAII	6.11E+01	8.87E+01	1.01E+02
	NSGAIII	4.35E+01	7.15E+01	8.00E+01
	MOEAD	1.51E+02	2.33E+02	3.54E+02
	SPEA2	3.34E+01	4.76E+01	6.15E+01
	GrEA	4.18E+01	6.37E+01	7.80E+01

Table 4. Comparisons of results of MFEA and other algorithms on 4-objective MOKP

Metric	Algorithm	Instance		
		4.250	4.500	4.750
		Mean	Mean	Mean
IGD	MFEA	3.87E+03	7.63E+03	1.01E+04
	NSGAI ^I	5.62E+03	1.09E+04	1.59E+04
	NSGAI ^{II}	6.32E+03	1.24E+04	1.85E+04
	MOEAD	5.28E+03	1.03E+04	1.41E+04
	SPEA2	5.70E+03	1.14E+04	1.67E+04
	GrEA	6.36E+03	1.27E+04	1.85E+04
HV	MFEA	2.80E+14	2.96E+15	9.16E+15
	NSGAI ^I	3.41E+14	4.15E+15	1.59E+16
	NSGAI ^{II}	2.32E+14	3.10E+15	1.30E+16
	MOEAD	5.72E+14	7.45E+15	3.09E+16
	SPEA2	3.29E+14	3.97E+15	1.55E+16
	GrEA	3.06E+14	4.03E+15	1.59E+16
CM	MFEA	1.61E+00	1.98E+00	1.96E+00
	NSGAI ^I	3.62E+00	4.20E+00	4.89E+00
	NSGAI ^{II}	8.66E+00	9.92E+00	1.34E+01
	MOEAD	2.46E+00	2.70E+00	2.53E+00
	SPEA2	4.01E+00	5.09E+00	5.76E+00
	GrEA	6.76E+00	8.13E+00	9.53E+00
GD	MFEA	3.87E+03	7.63E+03	1.01E+04
	NSGAI ^I	5.62E+03	1.09E+04	1.59E+04
	NSGAI ^{II}	6.32E+03	1.24E+04	1.85E+04
	MOEAD	5.28E+03	1.03E+04	1.41E+04
	SPEA2	5.70E+03	1.14E+04	1.67E+04
	GrEA	6.36E+03	1.27E+04	1.85E+04
SM	MFEA	2.01E+02	3.03E+02	3.96E+02
	NSGAI ^I	1.00E+02	1.81E+02	2.12E+02
	NSGAI ^{II}	5.91E+01	1.04E+02	1.27E+02
	MOEAD	2.19E+02	4.67E+02	6.68E+02
	SPEA2	5.90E+01	9.21E+01	1.14E+02
	GrEA	6.68E+01	1.08E+02	1.38E+02

From Fig. 5, it shows that MFEA identifies the true PF while others not, this phenomenon becomes more apparent as the number of items increases. The PF of MOEAD is far from the true. It can also be seen from Fig. 3, 4 that, the solutions of MFEA demonstrates the best distribution of solutions. The solutions of MFEA demonstrates well distribution, rapid convergence, and good diversity, presumably because that, MFEA involves a normalized mechanism, which integrates the inner and outer factor of each individual in the environment to overcome the conflicting nature of objectives. Equipped with Hybrid ageing, and Static hypermutation with FCM, MFEA effectively leverages upon the implicit parallelism of population-based search to exploit every corner of objective space, without bypassing or aggravating its conflicting nature.

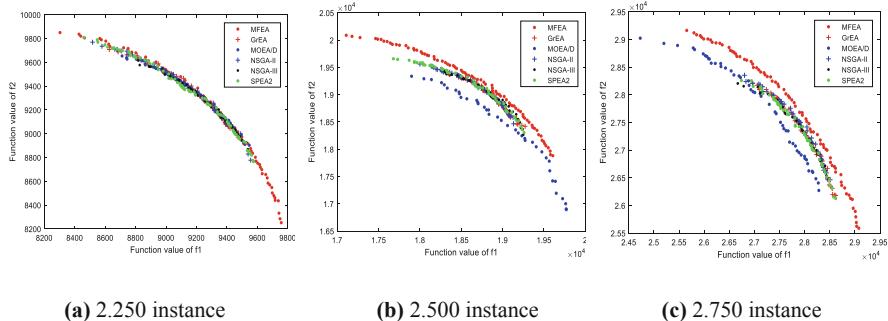


Fig. 5. Non-dominated fronts obtained for 2.250 instance.

5 Conclusion

In this paper, a novel MFEA is proposed. The effectiveness of MFEA is assessed on MOKP, compared with five state-of-the-art algorithms in terms of five performance metrics. Simulation results show that MFEA achieves the best performance in terms of both convergence and diversity. The proposed normalization mechanism is proved to be effective with nondominated-sorting, apparently ease the conflicted nature of objectives. Equipped with hybrid ageing operator and Static hypermutation with FCM, the drawback of MOEA are greatly improved. Overall, it is seen that this work is a significant contribution towards the understanding and application of the mixed-factor, nondominated-sorting, hybrid ageing operator, evolutionary algorithm and artificial immune system, i.e., hypermutation with FCM. It is reasonable that the future work will be focused on using this framework to solve engineering and real-life problems, such as scheduling, feature selection, location problem and other classical NP-hard combinatorial optimization problems, along with the theoretical insight into mixed factor and MFEA framework.

Acknowledgements. The authors would thank the reviewers for their valuable reviews and constructive comments. This work was supported by the project in the Education Department of Hainan Province (project number:Hnky2020–5), Hainan Provincial Natural Science Foundation of China (N0. 620QN237).

References

1. Schaffer, J.D.: Multiple objective optimization with vector evaluated genetic algorithms. In: Proceedings of the First International Conference on Genetic Algorithms and their Applications. Lawrence Erlbaum Associates. Inc. Publishers (1985)
2. Knowles, J.D., Corne, D.W.: Approximating the nondominated front using the pareto archived evolution strategy. *Evol. Comput.* **8**(2), 149–172 (2000)
3. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength pareto evolutionary algorithm. TIK-report, p. 103 (2001)
4. Zitzler, E., Künzli, S.: Indicator-Based Selection in Multiobjective Search. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiño, P., Kabán, A., Schwefel, H.-P. (eds.) Parallel Problem Solving from Nature - PPSN VIII. Lecture Notes in Computer Science, vol. 3242, pp. 832–842. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30217-9_84
5. Zhang, Q., Li, H.: MOEA/D: a multi-objective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **11**(6), 712–731 (2007)
6. Li, H., Deb, K., Zhang, Q., Suganthan, P.N., Chen, L.: Comparison between MOEA/D and NSGA-III on a set of many and multi-objective benchmark problems with challenging difficulties. *Swarm Evol. Comput.* **46**, 104–117 (2019). <https://doi.org/10.1016/j.swevo.2019.02.003>
7. Deb, K., Pratap, A., Agarwal, S., et al.: A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002)
8. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Trans. Evol. Comput.* **18**(4), 577–601 (2014)
9. Yang, S., Li, M., Liu, X., et al.: A grid-based evolutionary algorithm for many-objective optimization. *IEEE Trans. Evol. Comput.* **17**(5), 721–736 (2013)
10. Moslehi, F., Haeri, A.: An evolutionary computation-based approach for feature selection. *J. Ambient. Intell. Humaniz. Comput.* **11**(9), 3757–3769 (2019). <https://doi.org/10.1007/s12652-019-01570-1>
11. Ragmani, A., Elomri, A., Abghour, N., Moussaid, K., Rida, M.: FACO: a hybrid fuzzy ant colony optimization algorithm for virtual machine scheduling in high-performance cloud computing. *J. Ambient. Intell. Humaniz. Comput.* **11**(10), 3975–3987 (2019). <https://doi.org/10.1007/s12652-019-01631-5>
12. Wang, P., Xue, F., Li, H., Cui, Z., Xie, L., Chen, J.: A multi-objective DV-hop localization algorithm based on NSGA-II in internet of things. *Mathematics* **7**(2), 184 (2019). <https://doi.org/10.3390/math7020184>
13. Corus, D., Oliveto, P.S., Yazdani, D.: When hypermutations and ageing enable artificial immune systems to outperform evolutionary algorithm. *Theoret. Comput. Sci.* (2019). <https://doi.org/10.1016/j.tcs.2019.03.002>
14. Zhou, Z., Yang, Y., Qian, C.: Evolutionary Learning: Advances in Theories and Algorithms, pp. 6–9. Springer, Berlin (2019). <https://doi.org/10.1007/978-981-13-5956-9>
15. Lust, T., Teghem, J.: The multi-objective multidimensional knapsack problem: A survey and a new approach. *Int. Trans. Oper. Res.* **19**(4), 495–520 (2012)
16. Karp, R.M.: Reducibility among combinatorial problems. In: Miller, R.E., Thatcher, J.W., Bohlenger, J.D. (eds.) Complexity of Computer Computations, pp. 85–103. Springer US, Boston, MA (1972). https://doi.org/10.1007/978-1-4684-2001-2_9
17. Kumar, R., Banerjee, N.: Analysis of a multi-objective evolutionary algorithm on the 0–1 knapsack problem. *Theoret. Comput. Sci.* **358**(1), 104–120 (2006)

18. Zouache, D., Moussaoui, A., et al.: A cooperative swarm intelligence algorithm for multi-objective discrete optimization with application to the knapsack problem. *European J. Oper. Res.* **264**(1), 74–88 (2018)
19. Oliveto, P.S., Sudholt, D.: On the runtime analysis of stochastic ageing mechanisms. In: *Proceedings of the GECCO 2014*, pp. 113–120 (2014)
20. Zitzler, E., Thiele, L.: Multi-objective evolutionary algorithms: A comparative case study and the strength pareto approach. *IEEE Trans. Evol. Comput.* **3**(4), 257–271 (1999)
21. Van Veldhuizen, D.A., Lamont, G.B.: On measuring multi-objective evolutionary algorithm performance. In: *Proceedings of the 2000 Congress on Evolutionary Computation 2000*, pp. 204–211. IEEE (2000)
22. Zitzler, E., Thiele, L.: Multiobjective optimization using evolutionary algorithms—A comparative case study. In: Eiben, A.E., Bäck, T., Schoenauer, M., Schwefel, H.-P. (eds.) *Parallel Problem Solving from Nature—PPSN V*, pp. 292–301. Springer Berlin Heidelberg, Berlin, Heidelberg (1998). <https://doi.org/10.1007/BFb0056872>
23. Schott, J.R.: Fault tolerant design using single and multicriteria genetic algorithm optimization. AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH (1995)



NSLS with the Clustering-Based Entropy Selection for Many-Objective Optimization Problems

Zhaobin Ma¹, Bowen Ding¹, and Xin Zhang^{1,2(✉)}

¹ School of Artificial Intelligence and Computer Science, and Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214000, China
zhangxin@jiangnan.edu.cn

² Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

Abstract. The multi-objective optimization algorithm based on nondominated sorting and local search (NSLS) has shown great competitiveness in the most multi-objective optimization problems. NSLS can obtain the Pareto-optimal front with better distribution and convergence than other traditional multi-objective optimization algorithms. However, the performance of NSLS degrades rapidly when facing the many-objective optimization problems (MaOPs). This paper proposes another version of NSLS, named NSLS with the Clustering-based Entropy Selection (NSLS-CE), which replaces the farthest-candidate approach with the clustering-based entropy selection approach. The concept of clustering-based entropy is proposed to measure the distribution of populations, which is implemented by the k-means clustering algorithm. Besides, to reduce the time complexity of the proposed clustering-based entropy selection approach, we apply the thermodynamic component replacement strategy. In order to prove the efficacy of NSLS-CE for solving MaOPs, the experiment is carried out on eighteen instances with three different objective numbers. The experimental results indicate that NSLS-CE can obtain Pareto solutions with better convergence and better distribution than NSLS.

Keywords: Clustering-based entropy selection · Local search · Many-objective optimization problems

1 Introduction

Multi-objective optimization is a subfield of multi-criteria decision making, and it needs to make the optimal decision between two or more conflicting objectives. Multi-objective optimization has been widely applied in the fields of logistics, industry and finance [1–3]. For example, when completing the manufacturing tasks, companies need to consider various objectives including economic benefits, environmental protection, and social

benefits at the same time. Generally, a multi-objective optimization problem (MOP) can be formulated as:

$$\begin{aligned} \text{minimize : } F(\mathbf{x}) &= (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x}))^\top \\ \text{subject to : } \mathbf{x} &\in R^n \end{aligned} \quad (1)$$

where R^n is the decision variable space of n variables, $\mathbf{x} = [x_1, x_2, \dots, x_n] \in R^n$ is a solution vector, and $F : R^n \rightarrow R^m$ is constituted by m objective functions, R^m is the objective space.

Suppose that \mathbf{x}^a and \mathbf{x}^b are two solutions in the decision variable space, \mathbf{x}^a is said to dominate \mathbf{x}^b , if for each $i \in \{1, 2, \dots, m\}$ $f_i(\mathbf{x}^a) \leq f_i(\mathbf{x}^b)$ and at least one $j \in \{1, 2, \dots, m\}$ $f_j(\mathbf{x}^a) < f_j(\mathbf{x}^b)$. If \mathbf{x}^a does not dominate \mathbf{x}^a and \mathbf{x}^b does not dominate \mathbf{x}^a , \mathbf{x}^a and \mathbf{x}^b are said to be non-dominated. The Pareto-optimal solution is defined as that there is no other solution can dominate it. All the Pareto-optimal solutions in the search space form the Pareto-optimal set (PS). The Pareto-optimal front (PF) is composed of the objective vectors calculated by the solutions in PS.

NSLS is a promising algorithm for solving MOPs [4], which explores the solution space through the powerful search capability of the local search approach. Besides, the proposed farthest-candidate approach can maintain a well distribution of the population during the search process. NSLS is more competitive than the most multi-objective optimization algorithms in solving the problems with 2–3 objectives. However, with the increase of the number of objectives, MOPs become the many-objective optimization problems (MaOPs), the performance of NSLS decreases rapidly [5]. The reason is that the farthest-candidate approach is not suitable for the high-dimensional objective space. It cannot select the population with well distribution, the final population obtained by NSLS can only transmit in a small area of the real PF surface.

To solve MaOPs that NSLS fails on, this paper proposes an effective algorithm, NSLS with the clustering-based entropy selection, named NSLS-CE. NSLS-CE adopts the clustering-based entropy selection approach to replace the farthest-candidate approach. First, the selected population is classified into k clusters by the k-means clustering algorithm [6]. Then we propose the concept of clustering-based entropy, and select solutions from the k clusters based on the free energy minimization principle [7]. The experimental results demonstrate that the proposed algorithm NSLS-CE shows better competitiveness and performance on MaOPs.

2 Proposed Algorithm NSLS-CE

2.1 NSLS

NSLS, developed by Bili Chen, is a multi-objective optimization algorithm based on iterations. It applies local search to the field of multi-objective optimization. At each iteration, a better population P' is generated by the current population P through the proposed local search approach. Then the non-dominated sorting and farthest-candidate approach are applied to the combined population $P' \cup P$ to select a new population for the next iteration.

The local search approach of NSLS is shown as

$$n_{k,i}^+ = x_{k,i} + \lambda \times (x_{k,i}^1 - x_{k,i}^2) \quad (2)$$

$$n_{k,i}^- = x_{k,i} - \lambda \times (x_{k,i}^1 - x_{k,i}^2) \quad (3)$$

$$S_{k,i} = \{n_{k,i}^+, n_{k,i}^-\} \quad (4)$$

where $\mathbf{x}_i(x_{1,i}, \dots, x_{k,i}, \dots, x_{n,i})$, $k \in \{1, \dots, n\}$ is the i th solution of population P , $x_{k,i}$ is the k th variable of solution \mathbf{x}_i , $\mathbf{x}_i^1(x_{1,i}^1, \dots, x_{k,i}^1, \dots, x_{n,i}^1)$ and $\mathbf{x}_i^2(x_{1,i}^2, \dots, x_{k,i}^2, \dots, x_{n,i}^2)$ are two solutions randomly selected from population P , λ is a parameter obeying a Normal distribution with the mean value of μ and the standard deviation of σ , $n_{k,i}^+$ and $n_{k,i}^-$ represent the two neighbors of \mathbf{x}_i , $S_{k,i}$ are the set of neighbors of solution \mathbf{x}_i for the k th variable.

After generating $S_{k,i}$, a replacement strategy is adopted to decide whether the solutions in $S_{k,i}$ can replace solution \mathbf{x}_i . The replacement strategy is described as follows:

```

if  $n_{k,i}^+ \succ \mathbf{x}_i$  ( $n_{k,i}^+$  dominates  $\mathbf{x}_i$ ) and  $n_{k,i}^- \succ \mathbf{x}_i$  ,
    choose one to replace  $\mathbf{x}_i$  randomly;
else if  $n_{k,i}^+ \succ \mathbf{x}_i$  or  $n_{k,i}^- \succ \mathbf{x}_i$  ,
    choose the solution that dominates  $\mathbf{x}_i$  to replace  $\mathbf{x}_i$  ,
else if  $n_{k,i}^+ \not\sim \mathbf{x}_i$  ( $n_{k,i}^+$  and  $\mathbf{x}_i$  are non-dominated) and  $n_{k,i}^- \not\sim \mathbf{x}_i$  ,
    choose one to replace  $\mathbf{x}_i$  randomly;
else if  $n_{k,i}^+ \not\sim \mathbf{x}_i$  or  $n_{k,i}^- \not\sim \mathbf{x}_i$  ,
    choose the solution that is non-dominated with  $\mathbf{x}_i$  to replace  $\mathbf{x}_i$  ;
else
    do nothing;
end if

```

Algorithm 1 shows the pseudocode of the local search approach. NSLS performs well on the MOPs (the number of optimization objectives is two or three). However, as the number of the optimization problem objective increases, its performance drops dramatically. Hence, the proposed farthest-candidate approach fails in high-dimensional objective space, the well distributed solutions cannot be selected by this approach, resulting in the whole population loses good diversity in the search process.

To solve the problem that NSLS does not perform well in high-dimensional objective space, we proposed the clustering-based entropy selection approach to replace the farthest-candidate approach.

Algorithm 1 Local Search

Input: P : current population, N : number of population, n : number of variables

Output: P' : new population

```

1:  $P' = P;$ 
2: for  $i = 1$  to  $N$  do
3:   for  $k = 1$  to  $n$  do
4:     Calculate  $c = N(\mu, \sigma^2)$ ;
5:     Choose  $u_i$  and  $v_i$  from  $p$  randomly;
6:     Generate two neighborhood solutions  $w_{k,i}^+$  and  $w_{k,i}^-$  using (2) and (3),
      respectively;
7:     Replace  $x_i$  in  $p'$  according to the replacement strategy;
8:   end for
9: end for
10: return  $P'$ 
```

2.2 Clustering-Based Entropy Selection

Information entropy was defined by Shannon in 1948 [8], it is a very important concept in the field of information theory. Information entropy is related to thermodynamic entropy [9], it is used to describe the uncertainty about the occurrence of various possible events in the information source. Information entropy solves the measurement problem of information.

To measure the diversity of a population, we propose a variant of information entropy, named clustering-based entropy. First, the normalization operation is carried out on the objective function values of solutions in the population. The normalization formula is

$$f'_m(x) = \frac{f_m^{\min} - f_m(x)}{f_m^{\min} - f_m^{\max}} \quad (5)$$

where $f_m(x)$ is the m th objective function value of solution x , f_m^{\min} and f_m^{\max} are the minimum and maximum values of the m th objective function value in the population respectively.

We perform k-means clustering algorithm on the population according to the normalized objective function values after the normalization operation is done. The clustering-based entropy of population p is calculated as

$$H(p) = - \sum_{i=1}^k \frac{n_i}{N} \log_2 \frac{n_i}{N} \quad (6)$$

where k is the number of clusters, n_i is the number of solutions in the i th cluster, N is the number of solutions in population p .

The absolute energy of solution \mathbf{x} is defined as

$$e(i, \mathbf{x}) = \alpha \times (\frac{\pi}{2} - \theta_1) + (1 - \alpha) \times (\frac{\pi}{2} - \theta_2) \quad (7)$$

$$\theta_1 = \arccos \frac{c_i^T \cdot f'(\mathbf{x})}{\|c_i\| \times \|f'(\mathbf{x})\|} \quad (8)$$

$$\theta_2 = \min \left\{ \arccos \frac{f'(\mathbf{x}_\eta) \cdot f'(\mathbf{x})}{\|f'(\mathbf{x}_\eta)\| \times \|f'(\mathbf{x})\|} \mid \mathbf{x}_\eta \in p \text{ and } \mathbf{x}_\eta \neq \mathbf{x} \right\} \quad (9)$$

where c_i is the center point of the i th cluster, θ_1 is the angle between c_i and \mathbf{x} , θ_2 is the minimum angle value between solution \mathbf{x} and the other solutions in population p , and α is a number in $[0-1]$, in this paper α is equal to 0.25.

The relative energy of solution x is defined as

$$e'(i, x) = \frac{e(i, x) - e^{\min}}{e^{\min} - e^{\max}} \quad (10)$$

where e^{\min} and e^{\max} are the minimum absolute energy and maximum absolute energy in population p .

The free energy of population p can be expressed by

$$E(p) = \sum_{i=1}^k \sum_{x \in p_i} e'(i, x) \quad (11)$$

$$F(p) = E(p) - T \times H(p) \quad (12)$$

where p_i is the i th cluster of population p , $E(p)$ is the energy of population p , T is the temperature parameter.

According to Gibbs principle of minimum free energy, we need to select n solutions from population p , and minimize the free energy of the population composed of these n solutions. It is a difficult combinatorial optimization problem, and its time complexity is $O((n+k)C_N^n)$. To reduce the time complexity of this problem, the thermodynamic component replacement strategy is applied [10]. The thermodynamic component replacement strategy is described as follows:

- (1) Calculating the free energy component of each solution in population p , the free energy component of solution \mathbf{x} is shown as

$$F_c(i, \mathbf{x}) = e'(i, \mathbf{x}) + T \times \log_2 \frac{n_i}{N} \quad (13)$$

- (2) Selecting the n solutions with the smallest value of free energy component from population p .

Figure 1 illustrates a particular example of population selection by the clustering-based entropy selection approach. In this example, the population is divided into four

clusters. The white points are the solutions in the population, the white point connected to the broken line is the central point of each cluster, and the blue points are the solutions selected by the approach from the population. It is obvious that the points selected by the approach are distributed in the objective space uniformly. The pseudocode of clustering-based entropy selection is shown in Algorithm 2.

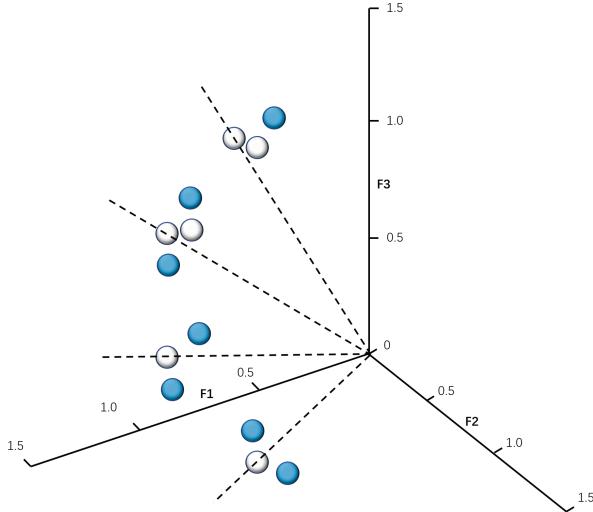


Fig. 1. Illustration of clustering-based entropy selection approach.

Algorithm 2 Clustering-Based Entropy Selection

Input: p : population, k : number of clusters, n : number of solutions selected from p , T : temperature parameter

Output: S : set of solutions

- 1: Normalize p using (5);
 - 2: Divide p into k clusters via k-means clustering algorithm;
 - 3: Compute the free energy component of each solution in p using (7)-(10) and (13);
 - 4: Select the first n smallest values of free energy component and put them into S ;
 - 5: **return** S
-

2.3 NSLS-CE

Algorithm 3 provide the pseudocode of the NSLS-CE algorithm. First, the population P with size N and other relevant parameters are initialized. The main loop of NSLS-CE algorithm is: first the current temperature parameter T is calculated, a better population P' is generated by the local search approach next. Then the current population P and the better population P' are combined as population U . After that, the nondominated fronts of U is generated by the fast nondominated sorting algorithm, and selecting solutions in nondominated fronts through lines 7–11 of Algorithm 3. Finally, the remaining solutions are selected by clustering-based entropy selection algorithm. Repeat the loop until the termination condition is met.

Algorithm 3 NSLS-CE

```

1: Initialize the values of  $P$ ,  $\mu$ ,  $\sigma$ ,  $t$ ,  $T_0$ , where  $P$  is the initial population with
size of  $N$  solutions, and it is created randomly.  $\mu$  and  $\sigma$  are the mean value
and the standard deviation of the Gaussian distribution respectively.  $t$  is
the generation number,  $T_0$  is the initial temperature;
2: while termination condition is not satisfied do
3:    $T = T_0/(1 + t)$ ;
4:    $P' = \text{LocalSearch}(P)$ ;
5:    $U = P' \cup P$ ;
6:   Generate all nondominated fronts of  $U: F = (F_1, F_2, \dots)$  by the fast non-
dominated sorting;
7:   Set  $P = \emptyset$  and  $i = 1$ ;
8:   while  $|P| + |F_i| \leq N$  do
9:      $P = P \cup F_i$ ;
10:     $i = i + 1$ ;
11:   end while
12:    $S = \text{Clustering-Based Entropy Selection}(F_i, \lceil \sqrt{|F_i|} \rceil, N - |P|, T)$ ;
13:    $P = P \cup S$ ;
14:    $t = t + 1$ ;
15: end while

```

3 Experimental Result and Discussion

3.1 Test Problems

In this paper, six test problems: DTLZ1–6 are employed to validate the performance of the NSLS-CE algorithm for solving MaOPs [11]. For each DTLZ test problem, the number of objectives is set as 7, 10, and 20.

3.2 Performance Measures

Performance metrics are used to evaluate the ability of evolutionary multi-objective optimization (EMO) algorithms to solve MaOPs. In this paper, we choose a widely used performance metric called inverse generational distance (IGD) metric.

IGD metric is a comprehensive performance evaluation metric [12]. IGD metric evaluates the convergence performance and distribution performance of an EMO algorithm by calculating the sum of the minimum Euclidean distances between each point on the real PF surface and the solutions in the PF obtained by the EMO algorithm. A smaller IGD indicates a better performance of convergence and distribution. IGD metric can be described as

$$IGD(S, P*) = \frac{\sum_{x \in P*} \min_{y \in S} dis(x, y)}{|P*|} \quad (14)$$

where $dis(x, y)$ is the Euclidean distance between point x and point y , S is a set of solutions obtained by an EMO algorithm, P^* is a set of points sampled evenly from the real PF surface.

3.3 Parameter Setting

The optimal parameter settings of NSLS-CE and NSLS are shown in Table 1. Both NSLS-CE and NSLS runs on the PlatEMO platform, and each test problem was tested with 20 replications for each algorithm.

Table 1. Optimal parameter settings of NSLS-CE and NSLS.

Algorithm	Tuned parameters
NSLS-CE	$N = 300$, $Max_FE = 60000$, $\mu = 0.5$, $\sigma = 0.1$, $T_0 = 5$, $\alpha = 0.25$
NSLS	$N = 300$, $Max_FE = 60000$, $\mu = 0.5$, $\sigma = 0.1$

3.4 Discussion of Results

The best, average and worst IGD metric values of NSLS-CE and NSLS on DTLZ1–6 are shown in Table 2. The best value among all the algorithms are in bold. Table 2 clearly indicates that NSLS-CE obtains better performance in most test problems.

In the test problem DTLZ3, as the number of objective increases to 10, the average IGD metric value obtained by NSLS is up to 69.8911. It indicates that the PF obtained by NSLS deviates from the real Pareto front surface, NSLS fails to maintain good distribution and convergence of the final population. The average IGD metric value obtained by NSLS-CE is 0.51099, which indicates that NSLS-CE can obtain the Pareto front surface with good convergence and distribution, and the Pareto front surface is close to the real Pareto front surface. NSLS only performs better than NSLS-CE on DTLZ2

(7M) and DTLZ4 (7M, 10M, 20M), and there is not much difference between these two algorithms. The boxplots of DTLZ1–6 (20M) are shown in Fig. 2, and it clearly suggests that NSLS-CE has the more consistent performance in all the test problems except DTLZ4.

Table 2. IGD metrics values comparison of NSLS-CE and NSLS on DTLZ1–6

Function	M		NSLS-CE	NSLS
DTLZ1	7	Best	0.09604	0.13843
		Avg	0.10797	0.31211
		Worst	0.11936	0.53945
	10	Best	0.12660	0.47848
		Avg	0.13833	1.02388
		Worst	0.15438	2.49920
	20	Best	0.22310	0.56906
		Avg	0.23184	2.06149
		Worst	0.24672	4.52880
DTLZ2	7	Best	0.28941	0.27875
		Avg	0.29252	0.28515
		Worst	0.29817	0.28986
	10	Best	0.42129	0.42239
		Avg	0.43197	0.43454
		Worst	0.43669	0.45380
	20	Best	0.67045	0.68392
		Avg	0.68471	0.70728
		Worst	0.69639	0.73253
DTLZ3	7	Best	0.31271	1.21410
		Avg	0.41645	3.93757
		Worst	1.11446	7.08680
	10	Best	0.47359	13.7260
		Avg	0.51099	69.8911
		Worst	0.56712	100.480
	20	Best	1.30420	87.9740
		Avg	2.07568	122.288
		Worst	3.24610	144.560

(continued)

Table 2. (*continued*)

Function	M		NSLS-CE	NSLS
DTLZ4	7	Best	0.38073	0.32909
		Avg	0.41959	0.36409
		Worst	0.45580	0.41026
	10	Best	0.57316	0.48746
		Avg	0.60478	0.49882
		Worst	0.62547	0.51513
	20	Best	0.71007	0.61674
		Avg	0.73140	0.62555
		Worst	0.74607	0.63798
DTLZ5	7	Best	0.06258	0.15199
		Avg	0.12653	0.18424
		Worst	0.20291	0.25684
	10	Best	0.08365	0.18900
		Avg	0.13881	0.29864
		Worst	0.18114	0.41819
	20	Best	0.19124	0.26126
		Avg	0.33815	0.38580
		Worst	0.49150	0.67688
DTLZ6	7	Best	0.03553	0.34242
		Avg	0.12393	0.57953
		Worst	0.21823	0.74209
	10	Best	0.18939	0.45021
		Avg	0.35198	0.78210
		Worst	0.74209	1.43410
	20	Best	0.19877	0.49205
		Avg	0.69217	4.06266
		Worst	1.67820	6.04570

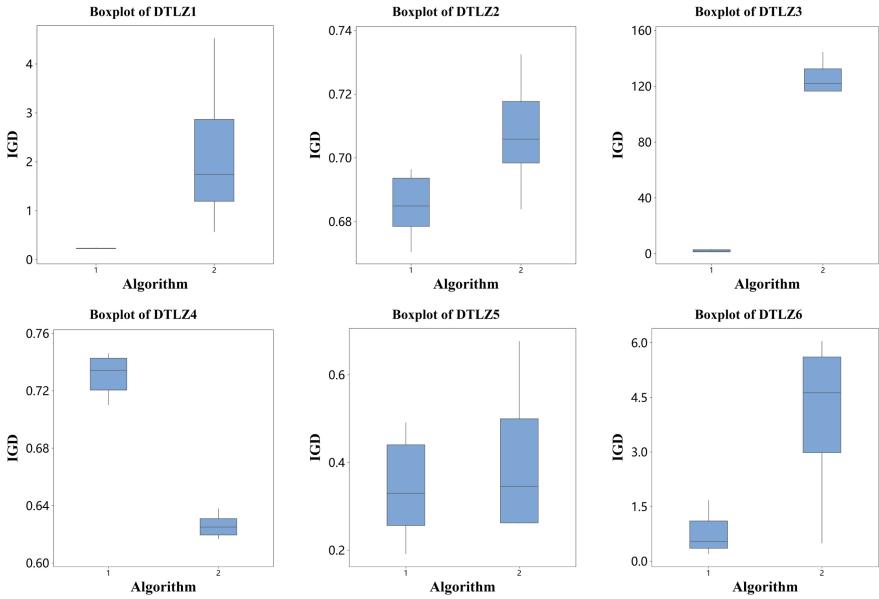


Fig. 2. Boxplots of algorithms on DTLZ1–6 (20M). Algorithm 1 and Algorithm 2 are NSLS-CE and NSLS respectively.

4 Conclusions

To address the rapid performance degradation of NSLS on MaOPs, this paper proposes an effective algorithm based on NSLS, named NSLS-CE. NSLS-CE improves the convergence and distribution of the population in the evolutionary process through the combination of local search approach and clustering-based entropy selection approach. In the clustering-based entropy selection approach, we propose the concept of clustering-based entropy, and implement it through the k-means clustering algorithm. Besides, the thermodynamic component replacement strategy is adopted to reduce the time complexity of the selection approach. This strategy can help to improve the accuracy of the selection while reducing the time complexity. Eighteen instances with three different optimization objective numbers are employed to verify the performance of the NSLS-CE algorithm. Experimental results show that NSLS-CE performs better than NSLS on the MaOPs. The proposed clustering-based entropy selection approach effectively solves the selection problem of NSLS on the high-dimensional objective problem. Moreover, the performance of NSLS-CE does not deteriorate as the number of objectives increases.

Acknowledgements. The work described in this paper was substantially supported in part by the National Natural Science Foundation of China under Grant No. 62106088, in part by the High-level personnel project of Jiangsu Province (JSSCBS20210852), and in part by the Fundamental Research Funds for the Central Universities, Jiangnan University (JUSRP121070), and Jilin University (93K172021K15).

References

1. Zhao, H., Chen, Z.G., Zhan, Z.H., et al.: Multiple populations co-evolutionary particle swarm optimization for multi-objective cardinality constrained portfolio optimization problem. *Neurocomputing* **430**, 58–70 (2021)
2. Wang, J., Ren, W., Zhang, Z., et al.: A hybrid multiobjective memetic algorithm for multiobjective periodic vehicle routing problem with time windows. *IEEE Trans. Syst. Man Cybern. Syst.* **50**(11), 4732–4745 (2018)
3. Ding, D., Zhang, X., Zhang, J., et al.: Multiobjective optimization of microwave circuits with many structural parameters and objectives. In: Proceedings of the 2019 International Conference on Microwave and Millimeter Wave Technology (ICMWT), pp. 1–3. IEEE (2019)
4. Chen, B., Zeng, W., Lin, Y., et al.: A new local search-based multiobjective optimization algorithm. *IEEE Trans. Evol. Comput.* **19**(1), 50–73 (2014)
5. Ishibuchi, H., Tsukamoto, N., Nojima, Y.: Evolutionary many-objective optimization: A short review. In: Proceedings of the 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), pp. 2419–2426. IEEE (2008)
6. Hartigan, J.A., Wong, M.A.: Algorithm AS 136 A k-means clustering algorithm. *J. Royal Stat. Soc. Series C (Appl. Stat.)* **28**(1), 100–108 (1979)
7. Callen, H.B.: Thermodynamics and an Introduction to Thermostatistics (1998)
8. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
9. Lieb, E.H., Yngvason, J.: The physics and mathematics of the second law of thermodynamics. *Phys. Rep.* **310**(1), 1–96 (1999)
10. Wei-Qin, Y., Yuan-Xiang, L., Cy, S.P.: Improving the computational efficiency of thermodynamical genetic algorithms (2008)
11. Deb, K., Thiele, L., Laumanns, M., et al.: Scalable multi-objective optimization test problems. In: Proceedings of the 2002 Congress on Evolutionary Computation. CEC 2002 (Cat. No. 02TH8600), vol. 1, pp. 825–830. IEEE (2002)
12. Sun, Y., Yen, G.G., Yi, Z.: IGD indicator-based evolutionary algorithm for many-objective optimization problems. *IEEE Trans. Evol. Comput.* **23**(2), 173–187 (2018)



Tunicate Swarm Algorithm Based Difference Variation Flower Pollination Algorithm

Chuchu Yu¹, Huajuan Huang^{2(✉)}, and Xiuxi Wei²

¹ College of Electronic Information, Guangxi Minzu University, Nanning 530000, China

² College of Artificial Intelligence, Guangxi Minzu University, Nanning 530000, China

hhj-025@163.com

Abstract. Flower pollination algorithm is a novel meta-heuristic swarm intelligence optimization algorithm, for its problems of insufficient solution accuracy, slow convergence speed and low stability, the Tunicate swarm algorithm based difference variation flower pollination algorithm (TSA-DVFPA) is proposed in this paper. The simplified Tunicate swarm algorithm and the random selection strategy were introduced into the process of cross-pollination. The differential variation strategy has been applied to the local search of the algorithm iteration to increase the diversity of the population. The improvement is validated by 16 benchmark functions. Compared with other similar algorithms, the results show that the proposed algorithm has a certain improvement in the convergence speed and the accuracy of optimization.

Keywords: Flower pollination algorithm · Tunicate swarm algorithm · Difference -variation · Function optimization

1 Introduction

The optimization problem can be traced back to the ancient extreme value problem, which has always been a key field of computer research [1]. In real life, many problems can be solved by traditional optimization models. However, there are also many practical problems that do not have a benign structure that need to be solved with the help of intelligent optimization algorithms. In recent years, meta-heuristic algorithms have attracted much attention, this type of algorithm which is constructed based on intuition and experience can solve complex combinatorial optimization problems that cannot be solved by traditional methods in a limited time. Inspired by the group behavior and evolutionary laws of natural creatures, there are many heuristic optimization algorithms have been proposed by some scholars, such as, particle swarm optimization (PSO) [2], sine cosine algorithm (SCA) [3], crow search algorithm (CSA) [4], Seagull optimization algorithm (SOA) [5], sparrow search algorithm (SSA) [6], and so on.

Flower pollination algorithm (FPA) [7] is a new meta-heuristic optimization algorithm that simulates the pollination behavior of flowering plants in nature and it was proposed by Cambridge University scholar Yang in 2012. From a biological point of view, pollination of plant flowers is a process of survival of the fittest and selection for

the best. The goal of pollination is to achieve the best reproduction of plants. FPA has the advantages of simple structure, few adjustment parameters, and certain stability. In addition, pollinators flight obeying the Levy distribution to make the global search more random and broaden the solution space. However, similar to other optimization algorithms, FPA also has some defects itself, such as slow convergence speed, insufficient convergence accuracy, easy trapping into the local optimum, and so on. Many scholars have made improvements to its shortcomings in recent years. In 2016, Zhou incorporated the Global elite opposition-based learning and the local self-adaptive greedy strategy into FPA to form a hybrid algorithm, which enhanced the algorithm's exploitation ability and the diversity of the population [8]. And then, Wang and Zhou proposed a variant of FPA to solve the Unmanned Undersea Vehicle path planning problem in two-dimensional and three-dimensional space. In the progress of iteration of this improved algorithm, a dimension by dimension-based update and evaluation strategy on solutions is used [9]. Part of scholars has combined FPA with another group of intelligent optimization algorithms to obtain an improved algorithm that has the advantages of both, such as, Emad Nabil [10] hybridized the standard FPA with the Clonal Selection Algorithm to obtain a modified flower pollination algorithm for global optimization in 2016. In 2018, Liu proposed a flower pollination algorithm based on sine cosine algorithm which means designing a sine cosine algorithm as a local optimization operator [11]. In 2020, Wang further simplified the sine cosine algorithm and incorporated it into FPA, and introduced the elite pollination operator to increase the stability of the algorithm and the precision of optimization [12]. Although the standard flower pollination algorithm has been researched by many scholars, there are still some problems in the improved algorithms, such as, the insufficient population caused by the small solution space of self-pollination which leads to the algorithm having a low optimization accuracy. On the contrary, the wide scope of the cross-pollination and the non-directional movement of pollens resulting in slow convergence of the algorithm.

In this paper, the tunicate swarm algorithm based differential variation flower pollination algorithm (TSA-DVFPA) is proposed. In the cross-pollination process, the simplified Tunicate Swarm Algorithm (TSA) [13] was introduced to improve the convergence speed of the original algorithm and reduce the blindness of the search process. In addition, the random selection operation was applied to avoid premature convergence. In the self-pollination stage, the differential variation strategy which selects basis vectors randomly was applied to ensure the population diversity and improve the optimization accuracy. And the dynamic switching probability strategy was introduced to balance the two pollination methods. In this paper, 16 benchmark functions were selected to experiment with the improved algorithm. Compared with six swarm intelligent optimization algorithms, the experimental results showed that TSA-DVFPA is feasible and effective.

The remainder of the paper is organized as follows: Sect. 2 briefly introduces the standard flower pollination algorithm; this is followed by the introduction of the algorithm proposed in the paper in Sect. 3; simulation experiments and results analysis are described in Sect. 4; Finally, conclusion and future works can be found in Sect. 5.

2 Flower Pollination Algorithm

Flower pollination algorithm is an optimization algorithm inspired by the pollination behavior of natural flowering plants. Based on the differences in pollination methods, the pollination process of flowering plants can be divided into two methods: cross-pollination and self-pollination. In order to simplify the pollination process, FPA obeys the following four idealized rules:

Rule 1: Cross-pollination behavior can be considered as that pollinators obey the Levy distribution to spread pollen globally.

Rule 2: The behavior of self-pollination can be considered as a local search step, and there are no biological pollinators.

Rule 3: Biological pollinators like bees can maintain the constancy of flowers, that is, the reproduction probability between pollinated flowers is proportional to their degree of similarity.

Rule 4: Due to the influence of many natural factors in the pollination process, the switching between cross-pollination and self-pollination is controlled by the switch probability p ($p \in [0, 1]$). The value of p has a profound impact on the efficiency of the algorithm.

The position X_i^t of pollen particle i for cross-pollination at iteration t is updated according to the following formula:

$$X_i^{t+1} = X_i^t + L(X_i^t - g_*) \quad (1)$$

where g_* is the optimal position among all pollens, and X_i^{t+1} represents solution vector X_i at iteration $t + 1$. Here L is a control parameter that represents the pollination intensity, which is essentially a random step size that obeys the Levy distribution. The distribution rule is shown in Eq. (2):

$$L \sim \frac{\lambda \Gamma(\lambda) \sin\left(\frac{\pi \lambda}{2}\right)}{\pi} \frac{1}{S^{1+\lambda}}, (s \gg s_0 \gg 0) \quad (2)$$

Here, $\Gamma(\lambda)$ is the standard gamma function.

At iteration i , the pollen particles self-pollinate according to the following equation:

$$X_i^{t+1} = X_i^t + \varepsilon (X_j^t + X_k^t) \quad (3)$$

where X_j^t and X_k^t are random individuals different from the population of current pollen. And ε is the proportional coefficient that obeys the uniform distribution in $[0, 1]$. According to the above analysis, the pseudo-code of FPA is as follows:

Algorithm 1. Flower pollination algorithm

Initialize a population of n flowers/pollen gametes with random solutions.

Find the best solution g^* in the initial population.

Define a switch probability $p \in [0,1]$.

Define a fixed number of iterations Max_iter .

While ($t < Max_iter$)

For $i = 1:n$

If $rand > p$

 Draw a step vector L which obeys a Levy distribution.

 Global pollination via Equation (1) and get a new solution X_i .

Else

 Draw ε from a uniform distribution in $(0,1)$.

 Do local pollination via Equation (3) and get new solution X_i .

End if

 Evaluate the new solutions.

 If new solutions are better, update them in the population.

End for

 Find the current best solution g_* .

End while

Output the best solution found.

3 Tunicate Swarm Algorithm Based Differential Variation Flower Pollination Algorithm

3.1 Strategy of Simplified TSA

Tunicate Swarm Algorithm (TSA) [13] is a new optimization algorithm proposed in 2020 which is inspired by the jet propulsion and swarm intelligence of tunicates in the foraging process. The vector \vec{A} is employed to calculating a new position of the search agent.

$$\vec{A} = \frac{\vec{G}}{\vec{M}} \quad (4)$$

Here, \vec{G} is the gravity force, \vec{M} represents the interaction between search agents. And the search agents move towards the optimal neighbor, following formula (5).

$$\vec{PD} = \left| \vec{FS} - r_{and} \times \vec{P_p(x)} \right| \quad (5)$$

where \vec{PD} indicates the distance between the food source and the searching individual, \vec{FS} indicates the location of the food source, $\vec{P_p(x)}$ is the position of the tunicate individual and the random number r_{and} is in the range $[0, 1]$.

In this paper, the effects of gravity \vec{G} and the interaction between search agents \vec{M} referred to in Eq. (4) are blurred and simplified into a random number ω that extended to indicate the influence of external factors on pollen pollinators. Due to the wide pollination scope in the cross-pollination step of FPA, which results in low optimization performance and slow convergence, the behavior of moving tunicate individuals toward the food source in TSA is integrated into FPA. The distance from the individual pollen to the optimal pollen is calculated as the vector of the direction in which the current pollen moves towards the optimal one.

$$\overrightarrow{PD} = |g_* - c \times X_i^t| \quad (6)$$

$$X_i^{t+1} = \begin{cases} g_* + \omega \times \overrightarrow{PD}, & \text{if } c \geq 0.5 \\ g_* - \omega \times \overrightarrow{PD}, & \text{if } c < 0.5 \end{cases} \quad (7)$$

Here, \overrightarrow{PD} is the distance vector from the current pollen to the optimal one, c and ω obey the uniform distribution in $(0, 1)$, g_* is the optimal position among all pollens.

The improvement strategy based on TSA undoubtedly enhances the algorithm's convergence ability, but it increases the risk of falling into a local optimum simultaneously. Therefore, the random selection strategy is added here to retain the random step of the pollen movement caused by Levy flight behavior:

$$x_i = \begin{cases} x_j, & \text{rand} > Cr \\ x_k, & \text{rand} \leq Cr \end{cases} \quad (8)$$

where Cr is considered as the choice rate, x_j and x_k are the positions of current pollen updated by the cross-pollination formula fused with TSA and the traditional cross-pollination formula, respectively.

3.2 Differential Variation Strategy of Local Pollination

The local search step of the basic flower pollination algorithm originally draws on the difference strategy in the standard differential evolution algorithm, and its update formula randomly selects the position vectors of two different pollens for the differential operation. The above strategy leads to an insufficient range of self-pollination, the optimization iteration only performs a small-scale search around the current optimal solution. Therefore, we proposed a differential variation strategy for self-pollination which can explore new solution sets within the current scope.

Differential variation operation achieves the purpose of enriching population diversity by randomly selecting basis vectors. The selection of the difference vector is improved as follows: the current pollen position is considered as the starting point of the vector, and a random vector different from the base one is selected as the endpoint. This improvement step reduces the randomness of the difference vector to a certain extent but maintains the randomness of the vector's endpoint. It prevents the whole iteration from being too greedy. And this strategy expands the range of r , and emphasizes the

influence of the difference vector to define the candidate solution. The updated formula of the variation vector is as follows:

$$X_i^{t+1} = X_j^t + r \times (X_k^t - X_i^t) \quad (9)$$

where X_j^t and X_k^t are positions of random pollens different from the current pollen particles at iteration t , and X_i^t is the position of current pollen in the solution space. Here r is a proportional coefficient that obeys uniform distribution in [1, 2].

3.3 Dynamic Switching Probability Strategy

In the FPA, the switch probability p is a fixed constant that controls the switching between global search and local search. However, in the later stage of the algorithm's evolution, the optimal individual has a strong attraction to others which greatly weakens the diversity of the population. That is, $|X_i^t - g_*|$ tends to 0. Therefore, the dynamic switching probability strategy is introduced to guarantee that the global search is emphasized in the early stage of the algorithm iteration process, and the local search is emphasized in the later stage. The following is the updated formula for switching probability p :

$$P = P_{min} + \frac{(P_{max} - P_{min}) \times iter}{Max_iter} \quad (10)$$

where P_{min} is assigned a value of 0 and the value of P_{max} is 0.8. Here, $iter$ represents the current iteration number, and Max_iter is the maximum of iterations.

3.4 Procedure of TSA-DVFPA

After the above three main improvements of FPA, the tunicate swarm algorithm based differential variant flower pollination algorithm (TSA-DVFPA) is obtained. The specific implementation steps are shown in Algorithm 2.

Algorithm 2. Tunicate swarm algorithm based differential variant flower pollination algorithm

Initialize a population of n flowers/pollen gametes with random solutions.
Find the best solution g_* in the initial population.
Define a parameter $Cr \in [0,1]$ for choice strategy.
Define a fixed number of iterations Max_iter .
While ($t < Max_iter$)
 Update the switch probability p via Equation (10).
 For $i = 1 : n$
 If $rand > p$
 If $rand > Cr$
 Update the distance between the best solution and search agent via Equation (6).
 Global pollination via Equation (7) and get new solution X_i .
 Swarm intelligent behavior.
 Else
 Draw a step vector L which obeys a Levy distribution.
 Global pollination via Equation (1) and get new solution X_i .
 End if
 Else
 Draw ε from a uniform distribution in (1,2).
 Do local pollination via Equation (9) and get a new solution X_i .
 End if
 Evaluate the new solutions.
 If new solutions are better, update them in the population.
End for
 Find the current best solution g_* .
End while
Output the best solution found.

3.5 Time Complexity Analysis of TSA-DVFPA

Assuming that the function of the optimization problem is $f(x)$ and the dimension of the solution space is D , then according to the description of the FPA algorithm and the operation rule of the time complexity symbol O , the time complexity of FPA [14] is $T(FPA) = O(D + f(D))$.

In TSA-DVFPA, n d -dimensional pollen particles are randomly generated in the initialization phase, and their time cost is $O(nD)$. In the process of K iterations, the time cost of global search and local search are both $O(K)O(nD)$, and the time complexity of dynamically generating switching probability p is $O(K)$. Since the random selection strategy can be generated by random functions with the same cost, its complexity is not shown. The time complexity T of TSA-DVFPA is:

$$T(TSA - DVFPA) = O(nD) + O(K)(O(nD) + O(nD) + O(1)) = O(D + f(D)).$$

Note that, the proposed algorithm has the same order of time complexity as the basic flower pollination algorithm.

4 Simulation Experiment and Results Analysis

This section is organized as follows: the experimental parameter settings is given in sect. 4.1 , and the comparison and analysis of experimental results can be found in Sect. 4.2. Finally, ablation experiments are described in Sect. 4.3.

4.1 Experimental Parameter Settings

Thirty independent experiments on three types of functions in Tables 2, 3, 4 were carried out with 7 algorithms. All comparison algorithms are as follows: Flower Pollination Algorithm (FPA), Mayfly Optimization Algorithm (MA) [15], Moth-Flame Optimization (MFO) [16], Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO) [17], Bat Algorithm (BA) [18].The population size is set to 100, and the maximum of iterations is set to 1000. The parameters of each algorithm are set in Table 1. The selected benchmark functions given in Table 2 are divided into three categories: high-dimensional unimodal functions ($f_{01} \sim f_{05}$), high-dimensional multimodal functions ($f_{06} \sim f_{09}$), and fixed-dimensional multimodal functions ($f_{10} \sim f_{16}$).

Table 1. The main parameter settings of the six algorithms.

Algorithms	Parameter values
TSA-DVFPA	$p \in [0, 0.8]$, $Cr = 0.3$, $r \in [1, 2]$
FPA	$p = 0.8$
MA	$r \in [-1, 1]$
MFO	$b = 1$, $r \in [-2, -1]$, $t \in [-1, 1]$
PSO	$w \in [0, 1]$, $c_1 = c_2 = 2$
GWO	$a \in [0, 2]$, $r_1 \in [0, 1]$, $r_2 \in [0, 1]$
BA	$F_{min} = 0$, $F_{max} = 1$, $A = 0.6$, $r = 0.7$

Table 2. Benchmark functions.

Benchmark Functions	Dim	Range	Optimum
$f_{01}(x) = \sum_{i=1}^n x_i^2$	30	$[-100, 100]$	0
$f_{02}(x) = \sum_{i=1}^n x_i + \prod_{i=1}^n x_i $	30	$[-10, 10]$	0
$f_{03}(x) = \max_i\{ x_i , 1 \leq i \leq D\}$	30	$[-100, 100]$	0
$f_{04}(x) = \sum_{i=1}^n (x_i + 0.5)^2$	30	$[-100, 100]$	0

(continued)

Table 2. (*continued*)

Benchmark Functions	Dim	Range	Optimum
$f_{05}(x) = \sum_{i=1}^n x_i^4 + \text{random}(0, 1)$	30	[-1.28, 1.28]	0
$f_{06}(x) = \sum_{i=1}^n \left[x_i^2 - 10\cos(2\pi x_i) + 10 \right]$	30	[-5.12, 5.12]	0
$f_{07}(x) = -20\exp\left(-0.2\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} - \exp\left(\frac{1}{n} \sum_{i=1}^n \cos 2\pi x_i\right)\right) + 20 + e$	30	[-32, 32]	0
$f_{08}(x) = \frac{1}{4000} \sum_{i=1}^n \left(x_i^2 \right) - \prod_{i=1}^n \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	30	[-600, 600]	0
$f_{09}(x) = \frac{\pi}{n} \left\{ 10\sin(\pi y_1) + \sum_{i=1}^{n-1} (y_i - 1)^2 \left[1 + 10\sin^2(\pi y_{i+1}) \right] + (y_n - 1)^2 \right\} + \sum_{i=1}^n \mu(x_i, 10, 100, 4)$	30	[-50, 50]	0
$f_{10}(x) = \sum_{i=1}^{11} \left[a_i - \frac{x_1(b_i^2 + b_i x_2)}{b_i^2 + b_i x_3 + x_4} \right]^2$	4	[-5, 5]	0.0003075
$f_{11}(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1 x_2 - 4x_2^2 + 4x_2^4$	2	[-5, 5]	-1.0316
$f_{12}(x) = \left(x_2 - \frac{5.1}{4\pi^2} x_1^2 + \frac{5}{\pi} x_1 + 6 \right)^2 + 10 \left(1 - \frac{1}{8\pi} \right) \cos x_1 + 10$	2	[-5, 10]	0.398
$f_{13}(x) = \left[1 + (x_1 + x_2 + 1)^2 \left(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1 x_2 + 3x_2^2 \right) \right] \times \left[30 + (2x_1 + 3x_2)^2 \left(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1 x_2 + 27x_2^2 \right) \right]$	2	[-2, 2]	3
$f_{14}(x) = - \sum_{i=1}^4 c_i \exp \left[\sum_{j=1}^3 a_{ij} (x_j - p_{ij})^2 \right]$	3	[0, 1]	-3.8628
$f_{15}(x) = - \sum_{i=1}^4 c_i \exp \left[\sum_{j=1}^3 a_{ij} (x_j - p_{ij})^2 \right]$	6	[0, 1]	-3.322
$f_{16}(x) = - \sum_{i=1}^5 \left[(x - a_i)(x - a_i)^T + c_i \right]^{-1}$	4	[0, 10]	-10.1532

4.2 Comparative Analysis of Benchmark Test Function Experiment Results

The test results of high-dimensional unimodal functions ($f_{01} \sim f_{05}$) are recorded in Table 3, while those of high-dimensional multimodal functions ($f_{06} \sim f_{09}$) and fixed-dimensional multimodal functions ($f_{10} \sim f_{16}$) are given in Table 4 and Table 5, respectively.

Table 3. Test results of high-dimensional unimodal functions.

Functions	Algorithms	Best	Worst	Mean	Std	Rank
f_{01}	TSA-DVFPA	0	0	0	0	1
	FPA	1.1106E+01	3.2267E+01	1.9999E+01	2.9544E+01	7
	MA	1.8632E-21	5.1191E-17	4.1104E-18	1.3192E-34	3
	MFO	1.7336E-06	1.0000E+04	6.6667E+02	6.2222E+06	4
	PSO	9.6681E-04	1.4600E-02	5.9000E-03	1.1190E-05	5
	GWO	1.0644E-87	1.4750E-84	1.5634E-85	7.767E-170	2
	BA	3.7194E+00	6.6553E+00	5.2489E+00	0.3708E+00	6
f_{02}	TSA-DVFPA	9.237E-203	6.171E-197	4.156E-198	0	1
	FPA	1.0602E+01	2.7585E+01	1.5012E+01	1.2228E+01	7
	MA	2.3333E-14	7.4028E-10	3.0800E-11	1.7609E-20	3
	MFO	1.7560E-01	9.0000E+01	3.0673E+01	3.8586E+02	5
	PSO	1.6080E-01	2.0640E+00	7.1110E-01	2.9770E-01	4
	GWO	2.8060E-50	2.5343E-48	2.9500E-49	2.0180E-97	2
	BA	9.0166E+00	8.5781E+02	5.3902E + 01	2.4523E+04	6
f_{03}	TSA-DVFPA	1.8755E-79	2.0290E-63	1.1020E-64	1.624E-127	1
	FPA	7.8188E+00	1.2454E+01	9.6522E + 00	1.3974E+00	5
	MA	5.8332E+00	2.8213E+01	1.5671E+01	3.0131E+01	6
	MFO	6.3556E+00	5.4139E+01	2.5742E+01	9.4914E+01	7
	PSO	1.7680E-01	1.7276E+00	6.2000E-01	1.0570E-01	3
	GWO	5.2403E-23	9.9069E-21	9.9079E-22	3.3233E-42	2
	BA	8.3730E-01	1.4069E+00	1.0734E+00	2.1800E-02	4
f_{04}	TSA-DVFPA	1.5000E-03	2.7100E-02	6.5000E-03	2.8127E-05	4
	FPA	1.1093E+01	3.3608E+01	1.9231E+01	3.6923E+01	7
	MA	1.2556E-21	1.3515E-17	8.0458E-19	5.9956E-36	1
	MFO	2.2237E-06	1.0100E+04	1.3467E+03	1.1788E+07	2
	PSO	2.6000E-03	1.5100E-02	7.1000E-03	1.3509E-05	5
	GWO	4.2930E-06	5.0230E-01	1.8200E-01	2.8900E-02	3
	BA	4.0625E+00	6.4674E+00	5.2411E + 00	2.8100E-01	6
f_{05}	TSA-DVFPA	4.4951E-06	9.4072E-04	1.6081E-04	2.9253E-08	1
	FPA	2.5900E-02	7.4500E-02	4.7700E-02	1.7518E-04	6
	MA	3.0000E-03	1.1300E-02	6.8000E-03	5.4254E-06	4
	MFO	2.0500E-02	1.6131E+01	1.4652E+00	1.3780E+01	5

(continued)

Table 3. (continued)

Functions	Algorithms	Best	Worst	Mean	Std	Rank
	PSO	1.7000E–03	5.3300E–02	1.6700E–02	1.1495E–04	3
	GWO	3.5574E–05	5.3354E–04	2.3635E–04	1.5300E–08	2
	BA	1.9412E+01	6.4311E+01	3.6665E+01	9.4391E+01	7

See from Table 3, in category 1, the performance of TSA-DVFPA is better than the other six algorithms for f_{01} , f_{02} , f_{03} and f_{05} . And for f_{01} and f_{02} , the standard deviation of TSA-DVFPA has reached 0. For f_{04} , the performance of TSA-DVFPA is not as good as MFO and GWO, but the optimal value searched by TSA-DVFPA is 4 orders of magnitude smaller than the optimal value obtained by FPA. In general, for high-dimensional unimodal functions, TSA-DVFPA has higher convergence accuracy and stronger robustness.

Similarly, as see in Table 4, for the high-dimensional multimodal functions, TSA-DVFPA can find the theoretical optimal values of f_{06} and f_{08} in each of 30 independent experiments, reflecting its strong stability. For function f_{09} , although the optimal solution of TSA-DVFPA is slightly worse than MFO and PSO, however, the precision of average fitness value and standard deviation of TSA-DVFPA were obtained from 30 experiments are better than those two algorithms. It can be seen that TSA-DVFPA has stronger stability than the two algorithms mentioned above. The experimental results show that TSA-DVFPA can effectively solve the problem of multimodal function in high-dimensional space.

In category three, we can easily discover that TSA-DVFPA can find the theoretical optimal value of all low-dimensional test functions from Table 5. And for f_{11} , f_{12} , f_{15} and f_{16} , the standard deviation obtained by the proposed algorithm is better than all comparison algorithms. For function f_{10} , although there are individual discrete values here that lower the average level of the algorithm, TSA-DVFPA can find the optimal solution in most independent experiments, and its performance is better than that of MFO, MA, and BA. Generally speaking, in the aspect of the optimization of fixed-dimensional multimodal functions, TSA-DVFPA has shown strong competition in optimization accuracy and optimization stability.

Table 4. Test results of high-dimensional multimodal functions.

Functions	Algorithms	Best	Worst	Mean	Std	Rank
f_{06}	TSA-DVFPA	0	0	0	0	1
	FPA	6.5196E+01	1.2599E+02	9.5522E+01	2.2418E+02	6
	MA	7.5724E–06	6.9763E+00	1.8581E+00	2.9565E+00	3
	MFO	6.3677E+01	2.2329E+02	1.3082E+02	1.7375E+03	5

(continued)

Table 4. (*continued*)

Functions	Algorithms	Best	Worst	Mean	Std	Rank
f_07	PSO	2.2615E+01	5.9547E+01	3.7136E+01	1.0196E+02	4
	GWO	0	1.1369E-13	3.7896E-15	4.1646E-28	2
	BA	1.9687E+02	2.9541E+02	2.5236E+02	5.7204E+02	7
f_08	TSA-DVFPA	8.8818E-16	8.8818E-16	8.8818E-16	0	1
	FPA	6.0197E+00	8.0712E+00	7.0583E+00	3.2540E-01	7
	MA	2.9766e-11	1.7068e-05	1.3707e-06	1.6656e-11	3
	MFO	5.7576E-04	1.9963E+01	6.5744E+00	8.6407E+01	4
	PSO	2.0229E+00	5.0291E+00	3.0567E + 00	5.9290E-01	5
	GWO	7.9936E-15	1.5099E-14	1.0125E-14	7.2365E-30	2
	BA	2.8904E+00	1.8709E+01	5.7284E + 00	2.9461E+01	6
f_09	TSA-DVFPA	0	0	0	0	1
	FPA	1.0811E+00	1.2504E+00	1.1651E + 00	1.5000E-03	6
	MA	0	7.5700E-02	1.8000E-02	5.0556E-04	3
	MFO	5.6497E-06	9.0745E+01	6.0441E+00	5.0886E+02	4
	PSO	1.4155E+00	7.3441E+00	3.3983E+00	2.8933E+00	7
	GWO	0	7.7000E-03	2.5767E-04	1.9254E-06	2
	BA	1.9720E-01	3.2430E-01	2.6250E-01	1.2000E-03	5

Table 5. Test results of Fixed-dimensional multimodal functions.

Functions	Algorithms	Best	Worst	Mean	Std	Rank
f_{10}	TSA-DVFPA	3.0750E-04	2.0400E-02	1.1000E-03	1.2917E-05	4
	FPA	3.0750E-04	3.0750E-04	3.0750E-04	4.6292E-23	1
	MA	3.0750E-04	2.0400E-02	2.2000E-03	3.0213E-05	5
	MFO	5.0293E-04	1.5000E-03	8.1298E-04	7.1786E-08	7

(continued)

Table 5. (*continued*)

Functions	Algorithms	Best	Worst	Mean	Std	Rank
	PSO	3.0750E–04	1.2000E–03	3.7632E–04	5.2071E–08	3
	GWO	3.0750E–04	1.2000E–03	3.3801E–04	2.7018E–08	2
	BA	3.9443E–04	1.6000E–03	8.7401E–04	6.2193E–08	6
f_{11}	TSA-DVFPA	−1.0316	−1.0316	−1.0316	0	1
	FPA	−1.0316	−1.0316	−1.0316	0	1
	MA	−1.0316	−1.0316	−1.0316	4.4373E–31	2
	MFO	−1.0316	−1.0316	−1.0316	0	1
	PSO	−1.0316	−1.0316	−1.0316	0	1
	GWO	−1.0316	−1.0316	−1.0316	2.2461E–18	3
	BA	−1.0316	−1.0306	−1.0314	3.8349E–08	4
f_{12}	TSA-DVFPA	3.9800E–01	3.9800E–01	3.9800E–01	0	1
	FPA	3.9800E–01	3.9800E–01	3.9800E–01	3.6519E–29	2
	MA	5.8444E+00	5.8444E+00	5.8444E+00	4.4702E–31	5
	MFO	3.9800E–01	3.9800E–01	3.9800E–01	0	1
	PSO	3.9800E–01	3.9800E–01	3.9800E–01	0	1
	GWO	3.9800E–01	3.9800E–01	3.9800E–01	2.6011E–14	3
	BA	3.9800E–01	3.9830E–01	3.9801E–01	4.4848E–09	4
f_{13}	TSA-DVFPA	3.0000E+00	3.0000E+00	3.0000E+00	8.8591E–28	1
	FPA	3.0000E+00	3.0052E+00	3.0007E+00	8.8487E–07	3
	MA	3.0000E+00	3.0000E+00	3.0000E+00	1.6237E–30	1
	MFO	3.0000E+00	3.0000E+00	3.0000E+00	7.3890E–30	1
	PSO	3.0000E+00	3.0000E+00	3.0000E+00	1.0249E–29	1
	GWO	3.0000E+00	3.0005E+00	3.0001E+00	1.3779E–08	2
	BA	3.0058E+00	3.3360E+00	3.1026E+00	5.8000E–03	4
f_{14}	TSA-DVFPA	−3.8628	−3.8628	−3.8628	8.8537E–22	1
	FPA	−3.8628	−3.8621	−3.8626	2.2852E–08	2
	MA	−3.8628	−3.8628	−3.8628	7.0997E–30	1
	MFO	−3.8628	−3.8628	−3.8628	3.0634E–30	1
	PSO	−3.8628	−3.8549	−3.8620	5.4891E–06	3
	GWO	−3.8628	−3.8549	−3.8608	7.6357E–06	4
	BA	−3.8529	−3.6898	−3.8009	1.8000E–03	5
f_{15}	TSA-DVFPA	−3.3220	−3.3220	−3.3220	1.3102E–22	1

(continued)

Table 5. (continued)

Functions	Algorithms	Best	Worst	Mean	Std	Rank
f_{16}	FPA	-3.3220	-3.2031	-3.2705	3.5000E-03	2
	MA	-3.3220	-3.2031	-3.2824	3.1000E-03	3
	MFO	-3.3220	-3.2031	-3.2348	2.8000E-03	5
	PSO	-3.3220	-2.8404	-3.2289	1.3800E-02	6
	GWO	-3.3220	-3.1377	-3.2597	4.0000E-03	4
	BA	-3.0372	-2.504	-2.7738	1.8900E-02	7
f ₁₆	TSA-DVFPA	-10.1532	-10.1532	-10.1532	4.1501E-20	1
	FPA	-10.1532	-5.0552	-6.2447	4.6493E+00	5
	MA	-10.1532	-2.6305	-5.9707	1.3486E+01	6
	MFO	-10.1532	-2.6305	-8.3206	9.5306E+00	3
	PSO	-10.1532	-2.6305	-7.5628	1.0644E+01	4
	GWO	-10.1532	-5.0552	-9.3078	3.5709E+00	2
	BA	-10.0102	-2.5544	-5.3993	9.0052E+00	7

The evolution curve diagrams of fitness values and the ANOVA test of the global minimum of some benchmark test functions are shown in Fig. 1, 2, 3, 4, 5, 6.

For the high-dimensional unimodal function f_{02} and the high-dimensional multimodal function f_{06} , the convergence curves Fig. 1 and Fig. 2 show that TSA-DVFPA converges the fastest, and the iteration curve in Fig. 2 shows that the improved algorithm has found the global optimum in less than 100 iterations. For the fixed-dimensional multimodal function f_{15} , TSA-DVFPA converges second only to GWO in the early iterations, and around 100 iterations, it converges faster than GWO and is the first algorithm to find the global optimum. In addition, as can be seen from Figs. 4, 5, 6, the variance test of TSA-DVFPA has no excess discrete values and a low median, reflecting the strong stability of the improved algorithm.

4.3 Ablation Experiments

This chapter conducts ablation experiments to verify the effectiveness of each improvement strategy. The three strategies proposed in Sect. 3.1, Sect. 3.2, and Sect. 3.3 are referred to here as strategy A, strategy B, and strategy C. The function f_{06} is used here as the test function. The three improvement strategies A, B, and C are removed respectively to form the comparison algorithm1 with improved strategies B and C, the comparison algorithm2 with improved strategies A and C, and the comparison algorithm3 with improved strategies A and B. Compare these three algorithms with the improved algorithm proposed in this paper, so as to verify the positive effect played by each improved strategy on the optimization results.

From the results of the ablation experiments (shown in Fig. 7), it can be seen that the biggest impact on the optimization results is strategy A proposed in the summary

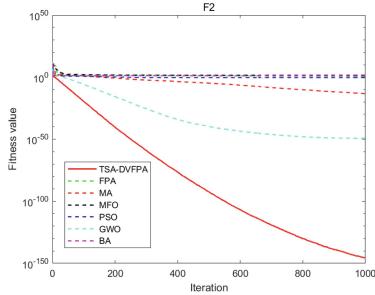


Fig. 1. Evolution curves of fitness value for f_{02}

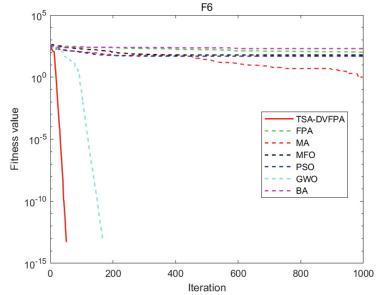


Fig. 2. Evolution curves of fitness value for f_{06}

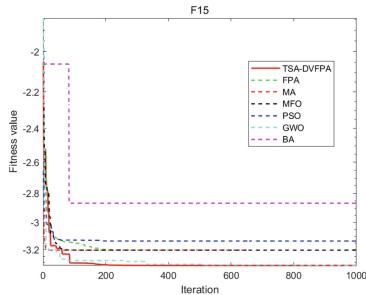


Fig. 3. Evolution curves of fitness value for f_{15}

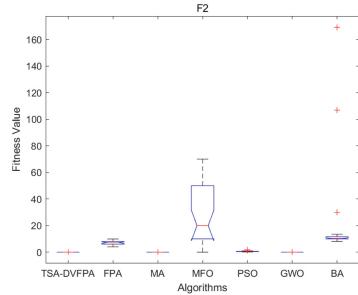


Fig. 4. ANOVA test of global minimum for f_{02}

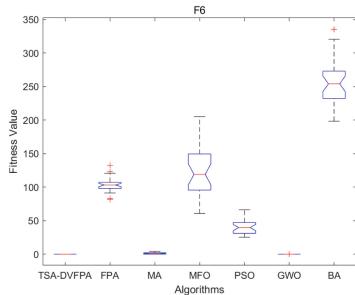


Fig. 5. ANOVA test of global minimum for f_{06}

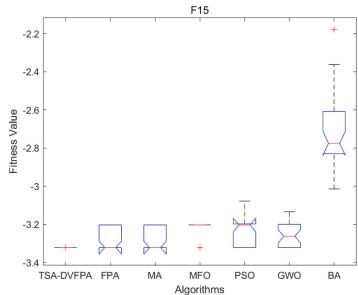


Fig. 6. ANOVA test of global minimum for f_{15}

of Sect. 3.1. At the same time, strategy B and strategy C can also effectively improve the convergence speed and optimization accuracy of the improved algorithm. In summary,

the three main improvement strategies have different degrees of positive impact on TSA-DVFPA. The improvement algorithm proposed in this paper does not depend on a single strategy but is the result of combining all the improvement ideas.

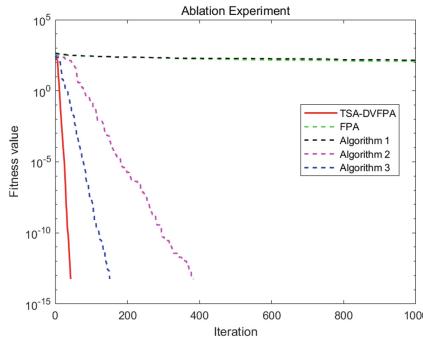


Fig. 7. Results of ablation experiments.

5 Conclusions and Future Works

In order to improve the slow convergence speed and low optimization accuracy of flower pollination algorithm, the tunicate swarm algorithm based differential variation flower pollination algorithm (TSA-DVFPA) is proposed in this paper. In the cross-pollination stage, inspired by the “jet propulsion” and “swarm behaviors” of tunicate swarms, the simplified tunicate swarm algorithm is introduced into the FPA. And the introduction of differential variation strategy in the self-pollination stage changes the original search boundary. From the experimental results, the overall performance of TSA-DVFPA is better than the basic flower pollination algorithm, it is also better than, or at least comparable with other comparison algorithms mentioned above. In general, the results show that the improved algorithm is feasible and effective and its convergence speed and accuracy of optimization have been improved to a certain extent.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (61662005). Guangxi Natural Science Foundation (2021GXNSFAA220068, 2018GXNSFAA294068).

References

1. Abhishek, K., Singh, R.P., Garcia, D.V., Rashmi, A.: *Swarm Intelligence Optimization: Algorithms and Applications*. John Wiley & Sons, Inc. (2020)
2. Kennedy, J.: Particle swarm optimization. *Encyclopedia of Machine Learning*. Springer, Boston, pp. 760–766 (2010)
3. Mirjalili, S.: SCA: a sine cosine algorithm for solving optimization problems. *Knowl. Based. Syst.* **96**, 120–133 (2016)

4. Askarzadeh, A.: A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm. *Comput Struct* **169**, 1–12 (2016)
5. Dhiman, G., Kumar, V.: Seagull optimization algorithm: theory and its applications for largescale industrial engineering problems. *Knowl Based Syst* **165**, 159–165 (2018)
6. Xue, J., Shen, B.: A novel swarm intelligence optimization approach: sparrow search algorithm. *Syst. ence Cont. Eng. An Open Access J* **8**(1), 22–34 (2020)
7. Yang, X.S.: Flower pollination algorithm for global optimization, In: Proceedings of the Unconventional Computation and Natural Computation, Lecture Notes Computer Science, vol. 7445, pp. 240–249 (2012)
8. Zhou, Y.Q., Wang, R., Luo, Q.: Elite opposition-based flower pollination algorithm. *Neurocomputing* **188**, 294–310 (2016)
9. Zhou, Y., Wang, R.: An improved flower pollination algorithm for optimal unmanned undersea vehicle path planning problem. *Int. J. Patt. Recog. Artif. Int.* **30**(4), 1–27 (2016)
10. Nabil, E.: A modified flower pollination algorithm for global optimization. *Expert Syst. Appl.* **57**, 192–203 (2016)
11. Liu, S., Zhao, Q.-H., Chen, S.-J.: Flower pollination algorithm based on sine cosine algorithm. *Microelectron. Comput.* **35**(06), 84–87 (2018)
12. Wang, L., Ding, Z.: Improved flower pollination algorithm combining sine cosine algorithm and elite operator. *Comput. Eng. Appl.* **56**(06), 159–164 (2020)
13. Kaur, S., Awasthi, L.K., Sangal, A.L., et al.: Tunicate swarm algorithm: a new bio-inspired based metaheuristic paradigm for global optimization. *Eng. Appl. Artif. Intell.* **2020**, 90 (2020)
14. Duan, Y., Xiao, H., Lin, Fang.: Flower pollination algorithm with new pollination methods. *Comput. Eng. Appl.* **54**(23), 94–108 (2018)
15. Zervoudakis, K., Tsafarakis, S.: A mayfly optimization algorithm. *Comput. Ind. Eng.* **145**, 106559 (2020)
16. Mirjalili, S.: Moth-flame optimization algorithm: a novel nature-inspired heuristic paradigm. *Knowl. Based. Syst.* **89**, 228–249 (2015)
17. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. *Adv. Eng. Softw.* **69**, 46–61 (2014)
18. Yang, X.S.: A new metaheuristic bat-inspired algorithm. *Nature Inspired Cooperative Strategies for Optimization* (NICSO 2010), Heidelberg, Springer, pp. 65–74 (2010)
19. Huang, P., Zheng, Q., Liang, C.: Overview of Image Segmentation Methods. *Wuhan Univ. (Nat. Sci. Ed.)* **66**(06), 519–531 (2020)



A Multi-strategy Improved Fireworks Optimization Algorithm

Pengcheng Zou¹, Huajuan Huang^{2(✉)}, and Xiuxi Wei²

¹ College of Electronic Information, Guangxi Minzu University, Nanning 530000, China

² College of Artificial Intelligence, Guangxi Minzu University, Nanning 530000, China

hhj-025@163.com

Abstract. To solve the shortcomings of traditional Fireworks Algorithm (FWA), such as slow convergence, being easy to fall into local optimum and low precision, a multi-operator improved Multi-strategy Fireworks Algorithm (MSFWA) was proposed. For initialization, the position of individual fireworks is initialized by chaos. As for the explosion operator, the explosion range is reduced nonlinearly and the explosion range of each fireworks particle is divided according to the level of fitness. It is beneficial to improve the development and exploration of the algorithm. For mutation operator, this paper adds mutation information on the basis of retaining the original information, and adopts adaptive strategy to select different mutation modes to further improve the ability to jump out of local optimum. For the selection operator, a brand-new strategy of multi-elite reservation + random / elite reservation is adopted, improving the global and local searching ability of the algorithm. Combining various strategies improves the global and local searching ability of the algorithm, and accelerates the convergence speed. Finally, 8 benchmark test functions and optimization problems of Design of Reducer are tested. The experimental results show that MSFWA has better optimization accuracy and performance than FWA and other heuristic intelligent algorithms.

Keywords: Fireworks algorithm · Multi-strategy · Self-adaptation · Dynamic selection · Engineering constrained optimization problem

1 Introduction

In recent years, meta-heuristic swarm intelligence algorithm has been paid more and more attention by scholars, and a series of swarm intelligence optimization algorithms have come out one after another, including Particle Swarm Optimization (PSO) [1], Fruit Fly Optimization Algorithm (FOA) [2], et al.

Fireworks Algorithm (FWA) [3] was proposed by Professor Tan Ying of Peking University in 2010. It is a swarm intelligence algorithm which simulates fireworks explosion and combines evolutionary computation with random search. Academic research on fireworks algorithm has been deepened and a series of improved algorithms have been proposed to address the shortcomings of the algorithm. Zheng, S.Q., Janecek, A. and

Tan, Y. [4] proposed an enhanced fireworks algorithm (EFWA) with minimum explosion radius threshold. The explosion radius calculated only based on the fitness difference of fireworks is difficult to reach a small value, so the local search ability of FWA and EFWA is poor. In view of the lack of local optimization ability, Zheng, S.Q., Janecek, A., Li, J., et al. [5] proposed a dynamic search fireworks algorithm (dynFWA) by introducing the explosion radius amplification and reduction mechanism. When the fireworks population searches for a position where the fitness value is no longer better, dynFWA will reduce the explosion radius of the core fireworks, and vice versa. Although dynFWA removes the mutation operator to accelerate the optimization speed, reducing the radius is easy to make the algorithm fall into local optimum when the fitness value is no longer good. Adaptive fireworks algorithm (AFWA) is another method proposed by Li, J., Zheng, S.Q. and Tan, Y. [6]. to adjust the search radius similar to the global search of particle swarm algorithm. The explosion radius of the current individual is the distance between the current individual and the current optimal individual so that the algorithm can jump out of the local optimum. Although AFWA can adaptively adjust the search step size and show good performance improvement, it is difficult to balance the mining and exploration ability of the algorithm with only the adaptive adjustment of step size and no adaptive change of the number of explosion sparks. In order to solve the above problems, this paper proposes a Multi-strategy Fireworks Algorithm (MSFWA). In this paper, the algorithm introduces chaos initialization, adaptive explosion radius and explosion number, adaptive switching of double mutation mode, multiple population elites + random selection and other strategies, and improves the initial population, explosion operation, mutation operation and selection operation of the original algorithm. 8 basic test functions are selected for testing and tested on pressure vessel engineering examples, which verifies the effectiveness and superiority of MSFWA algorithm in convergence ability and search ability. The remaining part of this paper is arranged as follows: the second part introduces the basic fireworks algorithm and the improved fireworks algorithm; the third part is the basic test function simulation and results analysis; the fourth part is the application of the improved algorithm in engineering examples; the fifth part is the summary and prospect of this paper.

2 Related Work

2.1 Fireworks Algorithm

FWA implementation steps are as follows:

Step 1: Initialize population. Randomly generate N fireworks. The fitness value of each firework is calculated according to the objective function, so as to determine the quality of fireworks.

Step 2: Calculate the explosion radius A_i of fireworks according to formulas (1), (2) and (3) and the number of explosion sparks S_i according to the different fitness values of fireworks. Different numbers of explosion sparks are produced under different explosion radii.

Step 3: The explosion spark is generated according to formula (4), and the explosion spark beyond the explosion space is detected and then is mapped into the feasible domain space according to formula (6).

Step 4: Generate a small amount of Gaussian mutation sparks according to formula (5), which increase population diversity. Also, the Gaussian mutation sparks generated by it are detected out of bounds, and the Gaussian mutation sparks beyond the explosion space are mapped into the feasible region space according to formula (6).

Step 5: Select the strategy. According to formula (7), $N - 1$ fireworks and the best contemporary fitness fireworks are selected as the next generation fireworks from the candidate set composed of fireworks, explosion sparks and Gaussian mutation sparks.

Step 6: Judge whether the termination condition is satisfied. If the condition is met, the search is terminated and the optimal value is output; Otherwise, go to Step 2.

Detailed parameters in formula 1–7 can be found in reference [3].

$$S_i = M \times \frac{y_{\max} - f(x_i) + \varepsilon}{\sum_{i=1}^N (y_{\max} - f(x_i)) + \varepsilon} \quad (1)$$

$$A_i = \hat{A} \times \frac{f(x_i) - y_{\min} + \varepsilon}{\sum_{i=1}^N (f(x_i) - y_{\min}) + \varepsilon} \quad (2)$$

$$S_i = \begin{cases} \text{round}(a * M), S_i < a * M \\ \text{round}(b * M), S_i \geq b * M \\ \text{round}(S_i), \text{else} \end{cases} \quad (3)$$

$$N_i^k = N_i^k + A_i \times U(-1, 1) \quad (4)$$

$$N_i^k = N_i^k \times e \quad (5)$$

$$N_i^k = N_{LB}^k + \text{mod}\left(\left|N_i^k\right|, \left(N_{UB}^k - N_{LB}^k\right)\right) \quad (6)$$

$$p_i = \frac{\sum_{j=1}^K d(N_i - N_j)}{\sum_{j=1}^K N_j} \quad (7)$$

2.2 Multi-strategy Fireworks Algorithm

Adaptive Explosion Strategy Step by Step. The fireworks population is divided into high quality, medium and poor grades according to the proportion R_1 , and the number of fireworks produced by the explosion of the three grades was denoted as Bn , Mn and Wn respectively. Excellent particles account for a small share, medium particles account for a large share, which helps the algorithm to give full play to its exploration ability at the initial stage of iteration. In the middle and late stage of iteration, high-quality particles gather with medium particles, and the fitness value is also small different. Some medium particles overlap with high-quality particles, but because of their different search radius, in the middle and late stage of iteration, the algorithm can search in a wide range and a small range near the current optimal solution position, which enhances the exploration

and production ability of the algorithm. The search step length greatly affects the search accuracy and efficiency of the algorithm. Therefore, an adaptive dynamic search radius is proposed in this paper. Zhang, S.P. and Wang, L.N. [7] applied Gaussian adaptive adjustment of search step length in FOA, which greatly improved the optimization accuracy of the algorithm. The mathematical expression is shown in formula (8). When it is applied to fireworks algorithm, the search step length is reduced very slowly at the initial stage of iteration, so that fireworks can fully explore the global optimal solution, and then quickly reduce the search step. After the middle and late iteration, the search step size began to decrease slowly to ensure the local optimization ability of the algorithm in the later stage.

$$fr = fr_M \times e^{-iter^2/(2 \times (MaxIter/\delta)^2)} \quad (8)$$

Detailed parameters in formula (8) can be found in reference [3]. To make different levels of fireworks give full play to their functions, this paper assigns different weights (ω_1) to different fireworks explosion radii according to the fitness value. The mathematical expression is shown in formula (9). When the algorithm falls into the local optimum, the search radius is still decreasing, so it fails to jump out of the local optimum. Therefore, an adaptive radius increasing coefficient (ω_2) is added in this paper. The mathematical expression is shown in formula (10).

$$\omega_1 = i \times (1 + C_1) \quad (9)$$

$$\omega_2 = 1 + C_2 \times \text{mod}(iter, m) \quad (10)$$

In the formula above, i represents the i th firework sorted by fitness, C_1 is a constant. C_2 is a constant. m represents increasing search radius for continuous $m - 1$ generation. The product of ω_1 , ω_2 and fr_M jointly determines the maximum value of the dynamic adaptive radius and the mathematical expression of adaptive dynamic radius in this paper is shown in formula (11).

$$A_i^{iter} = \omega_1 \times \omega_2 \times fr \quad (11)$$

Adaptive Optimization of Mutation Strategy. Selecting the dimension that needs to be mutated by mutation operation on some of its dimensions is thorny. Even if the dimension is selected, the direction and step length of the mutation will have a great impact. In short, the probability of the overall position becoming better is Relatively low. Then, Gaussian distribution characteristics determine that the algorithm only searches in the current solution domain, it is difficult to jump out of local optimum. To solve the first problem, the optimal selection method is adopted. The number of fireworks to be mutated is Mu , and the top $\text{ceil}(Mu/p)$ fireworks are ranked according to the fitness value. p adaptive mutation operations are performed on each firework to stop the operation until Mu fireworks are mutated. These fireworks are already in a relatively good position in the solution space, so it is beneficial to save the original good information and strengthen the local search ability by exploring them many times. In view of the second point, the adaptive mutation method is adopted. Cauchy mutation operator [8] has stronger global searching ability and enlarges the mutation range. The adaptive switching

coefficient α is introduced to realize the automatic switching of two mutation modes, and its mathematical expression is shown in formula (12) and the adaptive mutation operation is performed according to the formula (13). $e \sim N(0, 1), c \sim C(1, 0)$. l is the switching probability between Gaussian mutation and Cauchy mutation.

$$\alpha = iter / MaxIter \quad (12)$$

$$N_i^k = \begin{cases} N_i^k \times (1 + e), & \alpha > l \\ N_i^k \times (1 + c), & \alpha \leq l \end{cases} \quad (13)$$

‘Multi-elite + championship/elite’ strategy. ‘Multi-elite’ particles and ‘tournament / elite’ particles constitute candidate fireworks in proportion to R_2 . After a large number of experiments, the R_2 is set to 3:7, the process is as follows:

- a) 30% of the next generation’s fireworks N is occupied by ‘multi-elite’ particles, which refer to the elites of contemporary fireworks, explosive sparks and variant sparks that combine fine particles at 1:1:2.
- b) After removing the candidate ‘multi-elite’ particles from the original candidate set, the new candidate set KK is obtained by reordering according to the fitness value.
- c) ‘Championship / elite’ indicates the origin of 70% of the next generation of fireworks N according to the local optimal judgment criterion. If you fall into local optimum, use tournament selection method, otherwise use elite selection method.
- d) Championship selection: KK is approximately divided into three sets of KK_1 , KK_2 and KK_3 according to fitness. 70% of the number of fireworks is composed of the top-ranking particles in each set by 1:1:1.
- e) Elite choice: Select the first $ceil(0.7 \times N)$ particles directly from KK as the next generation fireworks.

The existence of ‘multi-elite’ in each generation not only ensures the diversity of fireworks population, but also improves the optimization ability of the algorithm because of its high-quality particles. The existence of ‘championship’ improves the algorithm’s ability to jump out of local optimum, while the existence of ‘elite’ further improves the algorithm’s optimization ability and accelerates the algorithm’s convergence speed. Through a large number of experiments, MSFWA has better performance than random selection strategy and roulette strategy by applying ‘multi-elite + championship / elite’ strategy.

Algorithm pseudo-code and its implementation.

Algorithm 1. MSFWA pseudo code

```

1 Input:  $N$  : population size,  $M$  : maximum number of explosion sparks,  $Mu$  : Variation
spark number;
2 Output: optimal solution and optimal fitness value;
3 Initialize  $N$  fireworks randomly in the search space;
4 Calculate the fitness value of fireworks;
5   For  $iter = 1$  to  $MaxIter$  do
6     For  $i = 1$  to  $N$  do
7       Calculate the number  $S_i$  of exploders according to the fitness level of fireworks ;
8       Calculate the explosion radius  $A_i^{iter}$  according to formulas (8), (9), (10) and (11) ;
9       According to the formula (4) to generate explosion sparks, and according to the formula
(6) for cross-border processing;
10      According to the formula (12), (13) to generate mutation spark, and according to the
formula (6) for cross-border processing;
11      Determine  $N$  fireworks as the next generation of fireworks;
12    End for
13  End for
14 Output the best solution and the best fitness;
```

Complexity Analysis of Algorithm. Because the practical application problem is usually more complex, it takes much time to calculate a particle fitness value ($O(f)$) than some simple operations (selection, judgment, assignment, etc.). According to the data in Table 2 and the analysis above, the sum of the total number of particles generated by FWA and MSFWA is shown in formula (14) and formula (15).

$$T_{FWA} = N + MaxIter \times (N + \sum_{i=1}^N S_i + Mu) \approx N + MaxIter \times (N + Mu + ceil(b \times M)) \quad (14)$$

$$\begin{aligned} T_{MSFWA} &= N + MaxIter \times (N + ceil(0.1 \times N \times (0.8 \\ &+ 0.7)/2 \times M + 0.75 \times N \times 0.5/2 \times M) + Mu) \\ &\approx N + MaxIter \times (N + Mu + ceil(0.26 \times N \times M)) \end{aligned} \quad (15)$$

Analysis shows that although b is usually less than $0.26 \times N$, for most complex problems, with the increase of N , $MaxIter$ usually decreases, so T_{MSFWA} is not much different from T_{FWA} , so MSFWA's fitness value calculates total consumption ($O(MSFWA) = T_{MSFWA} \times O(f)$) to the total FWA consumption ($O(FWA) = T_{FWA} \times O(f)$) similarly. In addition to the fitness value calculation consumption, the algorithm has some additional time overhead.

The additional time cost of FWA is mainly reflected in the explosion stage, variation stage and selection strategy stage, and the time complexity is $O(MaxIter \times \sum_{i=1}^N S_i \times d)$, $O(MaxIter \times Mu \times d)$ and $O(MaxIter \times (N + Mu + \sum_{i=1}^N S_i)^2 \times d)$, respectively.

The extra time cost of MSFWA is mainly reflected in the stage of explosion ($O(MaxIter \times ceil(0.26 \times N \times M) \times d)$), variation ($O(MaxIter \times Mu \times d)$) and selection ($O(MaxIter \times (N + Mu + ceil(0.26 \times N \times M)) \log_2(N + Mu + ceil(0.26 \times N \times M)) \times d)$).

Due to $O(n^2) >> O(n \log_2 n) >> O(n)$, the additional time overhead for FWA is denoted as $O(\text{MaxIter} \times (N + Mu + \sum_{i=1}^N S_i)^2 \times d)$ and then for MSFWA is denoted as $O(\text{MaxIter} \times (N + Mu + \text{ceil}(0.26 \times N \times M)) \log_2(N + Mu + \text{ceil}(0.26 \times N \times M)) \times d)$.

In summary, the time complexity of FWA algorithm is:

$$O(\text{FWA}) + O(\text{MaxIter} \times (N + Mu + \sum_{i=1}^N S_i)^2 \times d)$$

The time complexity of MSFWA algorithm is:

$$O(\text{MSFWA}) + O(\text{MaxIter} \times (N + Mu + \text{ceil}(0.26 \times N \times M)) \log_2(N + Mu + \text{ceil}(0.26 \times N \times M)) \times d)$$

For most problems, the total consumption ($O(\text{MSFWA})$) calculated by MSFWA fitness value is similar to the total consumption ($O(\text{FWA})$) of FWA, but $O(n^2) >> O(n \log_2 n) >> O(n)$, so the time complexity of MSFWA algorithm is usually less than that of FWA algorithm.

3 Experimental Simulation and Analysis

Due to too many parameters in this paper, the sensitivity of parameters is analyzed and the final parameters are determined by control variable experiment. In order to verify the effectiveness of the proposed MSFWA algorithm, it is used for simulation and comparative analysis in 8 standard test functions. The function names, dimensions, domains and optimal values of 8 test functions are listed in Table 1. In this paper, the simulation experiments are carried out from two aspects: (1) Compare MSFWA with standard FWA, GFWA [9] and LoTFWA [10], which have better performance in recent years, to test the overall improvement effect of the algorithm. (2) Compare MSFWA with the excellent swarm intelligence optimization algorithms of standard GWO [11], BOA [12] and WOA [13] in recent years to observe the overall performance of MSFWA algorithm in many algorithms.

3.1 Parameter Setting and Sensitivity Analysis

As for analyzing the parameter sensitivity, a large number of control variable experiments are carried out. Due to the limited length of the article, only some experimental comparisons are given. Table 2 shows the test results of the multimodal function in 30 independent experiments under different parameters. The parameter sensitivity is analyzed by the optimal value (best), average value (avg) and variance (std) of the test results.

It can be seen from Table 2 that although the change of some parameters makes the optimal value of the function better, it is often at the cost of the sacrifice of the average and variance, resulting in reduced algorithm stability. Considering various aspects, the final parameters are determined as shown in the first row of Table 2.

To reflect the fairness of the algorithm, the initial population number N in all the comparison algorithms is 30, and the specific parameter settings are shown in Table 3. In

the table, D is the search interval range. The maximum number of iterations is set to 1000. Then, the test is in 30 independent experiments. The experiments of all algorithms run on the same platform, and the experimental platform is MATLAB R2018b under Windows 10(64bit) operating system. Hardware conditions are Intel (R) core (TM) i5-4210 CPU 2.90GHz and 8GB memory.

3.2 Experimental Results and Analysis

In order to test the overall improvement effect of MSFWA algorithm, Table 4 lists the detailed data of the optimal value, average value, standard deviation and iteration time of objective function, in which the best search results are marked in bold. To make a more intuitive comparison, the search quality is quantified in the table, sorted by mean and variance, and the higher the ranking, the smaller the grade value. It can be seen that except for the algorithm ranked first in F_7 , MSFWA is superior to the other six algorithms in other test functions. According to the ‘no free lunch’ theorem, no single algorithm can take all the advantages. Although MSFWA ranks second in F_7 function, MSFWA has advantages in solving unimodal function problems, and it is not easy to fall into local optimum when solving multimodal functions.

To supplement the explanation, Figure 1 to 2 describe the iterative comparison curves of F_3 and F_4 functions. Figure 3 to 4 respectively describe the variance diagrams of the 30 times optimal values of F_6 and F_8 functions, where the abscissa is the algorithm name and the ordinate is the function value. It is observed from the diagram that MSFWA can find the optimal value earlier with the same average accuracy (F_4). Under the conditions of unimodal function (F_2 , F_3) and multimodal function (F_4), the slope of the function convergence curve is the largest, that is, the convergence speed is the fastest, and the solution with higher accuracy can be found. In the multimodal function (F_4), MSFWA is not easy to ‘premature’ and quickly jumps out of local extremum. In the environment of function F_1 , F_6 and F_8 , the optimal value of MSFWA is the best and has the smallest fluctuation compared with other comparison algorithms. The experimental data, convergence diagram and variance diagram show that MSFWA algorithm performs better than other FWA variants and other intelligent optimization algorithms in the standard function test.

3.3 Statistical Test

In the comparison of the optimization performance of each algorithm, it is not perfect to take the optimal value, average value and standard deviation as the evaluation basis for the effectiveness of the algorithm. In order to further compare the optimization effect of each algorithm, Wilcoxon rank sum test with the significance level of 0.05 is performed on the test results of each algorithm based on Table 3. MSFWA is compared with FWA, GFWA, LotFWA, GWO, BOA and WOA on eight different test functions one by one. Table 5 is the p value of rank sum test of MSFWA and the optimization results of each algorithm. When the p value is less than 0.05, it indicates that the optimization effects of the two algorithms are different, and vice versa. ‘NaN’ means that all contrast algorithms find the global optimal value.

It can be seen from Table 5 that MSFWA and GFWA in F_7 function and MSFWA and GFWA and LoTFWA in F_8 function have certain similarities. In other functions, MSFWA is very different from other comparison algorithms, indicating that MSFWA has good optimization performance.

4 Application of MSFWA in Engineering Constrained Optimization Problem

In order to verify the performance of MSFWA in optimizing practical engineering problems, this algorithm and FWA are applied to the optimization problem of reducer design [14]. The objective function and constraint conditions of the problem can be found in reference [15]. The algorithm runs independently for 30 times, and other parameter settings are shown in Table 3.

Compared with FWA, differential evolution with dynamic random selection (DEDS) [17], artificial bee colony algorithm (ABC) [16], neighborhood adaptive constrained fractional particle swarm (NAFPSO) [17] and horizontal comparative differential evolution (DELc) [18], the test results are shown in Table 6. Compared with other algorithms, the MSFWA has achieved good results in Design of Reducer. In the aspect of obtaining the optimal solution, MSFWA obtained the optimal value of 2896.2953, which is superior to other comparison algorithms, reflecting its excellent optimization ability. In the aspect of algorithm stability, the MSFWA optimization variance is much smaller than FWA, and slightly inferior to other comparison algorithms. Generally speaking, MSFWA is superior to the considered method in robustness and effectiveness. To supplement the improved performance, Fig. 5 and Fig. 6 are MSFWA and FWA convergence diagram and variance diagram of Design of Reducer. It can be seen that the convergence speed, convergence accuracy and stability of MSFWA are better than those of FWA, showing the excellent effect of the improved MSFWA.

5 Concluding Remarks

Aiming at the shortcomings of FWA algorithm, this paper improves the traditional fireworks algorithm, introduces Gaussian adaptive search step strategy, adaptive mutation strategy with double mutation mode switching and ‘multi-elite + championship / elite’ strategy into the global optimization and local optimization mechanism of fireworks algorithm, and proposes a multi-strategy fireworks algorithm. Through the performance test of MSFWA algorithm in high dimension and low dimension and the test of reducer design engineering constraint optimization problem, the results show that the MSFWA algorithm has stronger competitive advantage than other algorithms. In future work, we can consider adding the evolutionary mechanism of more advanced algorithms and multi-strategy operators to fireworks algorithm, and apply MSFWA to more complex optimization problems and practical applications.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (61662005). Guangxi Natural Science Foundation (2021GXNSFAA220068, 2018GXNSFAA294068).

Appendix

Table 1. Standard test function

Function	Function name	Domains	Dimensions	Optimal Value
F ₁	Rosenbrock	[−30,30]	30	0
F ₂	Sphere	[−100,100]	30	0
F ₃	Zakharov	[−5, 10]	30	0
F ₄	Griewank	[−600,600]	30	0
F ₅	Rastrigin	[−5.12,5.12]	30	0
F ₆	Penalized	[−50,50]	30	0
F ₇	Shubert	[−5.12,5.12]	2	−186.7309
F ₈	Shekel	[0,10]	4	−10.5364

Table 2. Test function of F6

R ₁	Bn (M)	Mn (M)	Wn (M)	C1	C2	m	fr _M (D)	p	l	R ₂	best	avg	std
2:15:3	0.7	0.5	0	0.3	0.05	8	0.05	3	0.3	3:7	2.59E-16	7.32E-04	2.74E-03
1:2:1	0.7	0.5	0	0.3	0.05	8	0.05	3	0.3	3:7	5.45E-15	3.16E-04	8.13E-03
2:15:3	0.7	0.3	0.1	0.3	0.05	8	0.05	3	0.3	3:7	2.40E-12	2.20E-03	4.40E-03
2:15:3	0.7	0.5	0	0.8	0.05	8	0.05	3	0.3	3:7	2.09E-19	1.60E-03	3.90E-03
2:15:3	0.7	0.5	0	0.3	0.3	8	0.05	3	0.3	3:7	4.10E-15	9.49E-04	2.40E-03
2:15:3	0.7	0.5	0	0.3	0.05	2	0.05	3	0.3	3:7	1.05E-18	1.10E-03	3.30E-03
2:15:3	0.7	0.5	0	0.3	0.05	8	0.2	3	0.3	3:7	6.73E-15	1.10E-03	3.30E-03
2:15:3	0.7	0.5	0	0.3	0.05	8	0.05	1	0.3	3:7	1.92E-19	1.10E-03	3.30E-03
2:15:3	0.7	0.5	0	0.3	0.05	8	0.05	3	0.1	3:7	2.23E-18	1.10E-03	3.30E-03
2:15:3	0.7	0.5	0	0.3	0.05	8	0.05	3	0.3	1:1	2.77E-15	2.20 E-03	4.36E-03

Table 3. Specific parameters

Algorithm	Parameters
FWA	$N = 30, M = 20, Mu = 10, a = 0.04, b = 0.8, \hat{A} = 0.8D$
GFWA	$N = 30, M = 20, \sigma = 0.2, \alpha = 0, A_c(1) = 0.8D, \rho_+ = 1.2, \rho_- = 0.9$
LoTFWA	$N = 30, M = 20, \sigma = 0.2, A_c(1) = 0.8D, \rho_+ = 1.2, \rho_- = 0.9$

(continued)

Table 3. (*continued*)

Algorithm	Parameters
MSFWA	$N = 30, M = 20, Mu = 10, fr_M = 0.05D, \delta = 6, p = 3, C_1 = 0.3, C_2 = 0.05, m = 8, l = 0.3$
GWO	$N = 30$
BOA	$N = 30, p = 0.8, c = 0.01, a = 0.3$
WOA	$N = 30$

Table 4. Test results of standard function

Function	algorithm	average value	variance	optimal value	Time/sec	ranking
F ₁	MSFWA	5.77E-04	8.40E-04	5.59E-08	2.31E + 00	1
	FWA	2.87E + 01	1.85E-01	2.81E + 01	3.65E + 00	6
	GFWA	1.05E + 01	1.28E + 01	6.06E-04	1.21E + 01	3
	LoTFWA	2.54E + 01	1.70E + 00	2.15E + 01	8.81E + 00	4
	GWO	2.68E + 01	7.12E-01	2.51E + 01	7.49E-01	5
	BOA	2.90E + 01	7.37E-03	2.90E + 01	1.21E + 00	7
	WOA	2.18E + 01	1.09E + 01	2.83E-04	5.59E-01	2
F ₂	MSFWA	0.00E + 00	0.00E + 00	0.00E + 00	2.97E + 00	1
	FWA	3.43E-86	1.61E-85	1.09E-121	3.51E + 00	3
	GFWA	1.49E-11	3.08E-11	1.56E-14	1.10E + 01	7
	LoTFWA	6.43E-15	1.03E-14	1.61E-16	8.09E + 00	6
	GWO	4.77E-59	1.02E-58	8.07E-61	7.06E-01	4
	BOA	7.15E-23	1.68E-23	4.80E-23	1.10E + 00	5
	WOA	7.56E-149	3.96E-148	2.34E-171	5.14E-01	2
F ₃	MSFWA	0.00E + 00	0.00E + 00	0.00E + 00	2.68E + 00	1
	FWA	3.57E-59	1.92E-58	5.35E-88	3.30E + 00	2
	GFWA	4.07E-09	4.65E-09	4.28E-11	1.07E + 01	5
	LoTFWA	7.68E-05	7.49E-05	5.35E-06	8.22E + 00	6
	GWO	2.47E-19	3.73E-19	4.44E-22	7.18E-01	4
	BOA	7.12E-23	2.11E-23	4.01E-23	1.15E + 00	3
	WOA	1.54E + 02	2.56E + 02	9.82E-06	5.17E-01	7

(continued)

Table 4. (*continued*)

Function	algorithm	average value	variance	optimal value	Time/sec	ranking
F ₄	MSFWA	0.00E + 00	0.00E + 00	0.00E + 00	5.13E + 00	1
	FWA	0.00E + 00	0.00E + 00	0.00E + 00	4.07E + 00	1
	GFWA	3.04E-03	6.59E-03	2.95E-14	1.39E + 01	5
	LoTFWA	6.89E-03	9.96E-03	1.78E-15	1.03E + 01	7
	GWO	5.04E-03	1.19E-02	0.00E + 00	7.88E-01	6
	BOA	5.07E-16	1.27E-16	3.33E-16	1.31E + 00	4
	WOA	0.00E + 00	0.00E + 00	0.00E + 00	5.72E-01	1
F ₅	MSFWA	0.00E + 00	0.00E + 00	0.00E + 00	4.29E + 00	1
	FWA	0.00E + 00	0.00E + 00	0.00E + 00	4.20E + 00	1
	GFWA	3.24E + 01	2.67E + 01	1.29E-09	1.39E + 01	6
	LoTFWA	1.34E + 02	3.05E + 01	6.67E + 01	9.88E + 00	7
	GWO	7.71E-01	2.03E + 00	0.00E + 00	8.35E-01	5
	BOA	3.30E-13	1.48E-13	1.14E-13	1.37E + 00	4
	WOA	0.00E + 00	0.00E + 00	0.00E + 00	6.03E-01	1
F ₆	MSFWA	7.32E-04	2.74E-03	2.59E-16	3.16E + 00	1
	FWA	2.82E + 00	2.50E-01	1.89E + 00	3.73E + 00	6
	GFWA	5.40E-03	7.16E-03	1.84E-13	1.83E + 01	2
	LoTFWA	5.85E-03	1.05E-02	3.39E-12	1.23E + 01	3
	GWO	4.99E-01	2.35E-01	3.32E-05	9.20E-01	5
	BOA	3.00E + 00	3.77E-04	3.00E + 00	1.85E + 00	7
	WOA	9.15E-02	1.34E-01	8.78E-11	7.32E-01	4
F ₇	MSFWA	-1.8673E + 02	1.7976E-14	-1.8673E + 02	2.37E + 00	2
	FWA	-1.8670E + 02	4.3364E-02	-1.8673E + 02	2.78E + 00	5
	GFWA	-1.8673E + 02	2.4886E-14	-1.8673E + 02	1.09E + 01	3
	LoTFWA	-1.8673E + 02	0.0000E + 00	-1.8673E + 02	8.00E + 00	1
	GWO	-1.8463E + 02	1.1337E + 01	-1.8673E + 02	5.04E-01	6
	BOA	-1.7195E + 02	1.5452E + 01	-1.8671E + 02	1.16E + 00	7
	WOA	-1.8673E + 02	2.2126E-05	-1.8673E + 02	5.10E-01	4
F ₈	MSFWA	-1.0536E + 01	2.6348E-15	-1.0536E + 01	2.19E + 00	1
	FWA	-9.1451E + 00	2.1232E + 00	-1.0498E + 01	3.10E + 00	7

(continued)

Table 4. (*continued*)

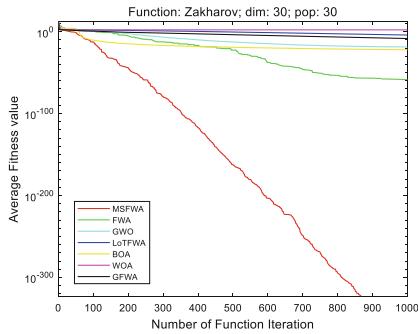
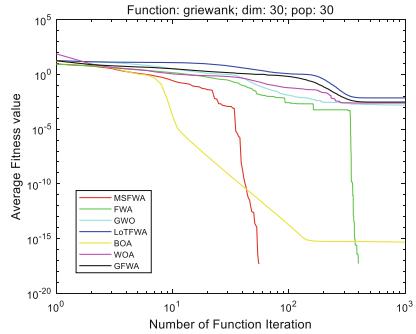
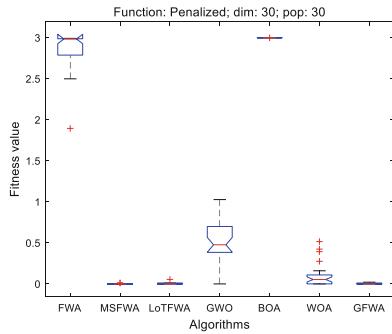
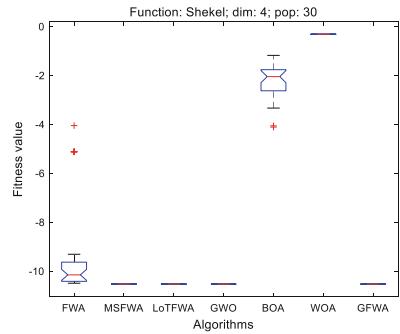
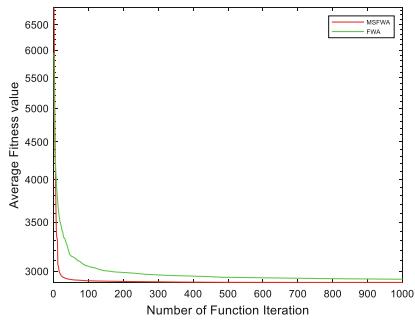
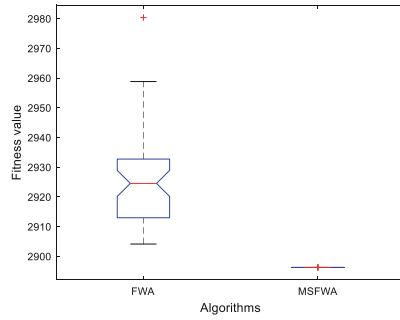
Function	algorithm	average value	variance	optimal value	Time/sec	ranking
	GFWA	-1.0536E + 01	3.4270E-15	-1.0536E + 01	1.00E + 01	2
	LoTFWA	-1.0536E + 01	4.9399E-15	-1.0536E + 01	7.72E + 00	3
	GWO	-1.0536E + 01	2.4977E-04	-1.0536E + 01	5.09E-01	4
	BOA	-2.2746E + 00	7.4675E-01	-4.1159E + 00	1.07E + 00	5
	WOA	-3.2173E-01	0.0000E + 00	-3.2173E-01	4.84E-01	6

Table 5. Wilcoxon rank sum test results

Function	FWA	GFWA	LoTFWA	GWO	BOA	WOA
F ₁	1.83E-04	1.01E-03	1.83E-04	1.83E-04	1.83E-04	1.83E-04
F ₂	6.39E-05	6.39E-05	6.39E-05	6.39E-05	6.39E-05	6.39E-05
F ₃	6.39E-05	6.39E-05	6.39E-05	6.39E-05	6.39E-05	6.39E-05
F ₄	NaN	6.39E-05	2.10E-03	NaN	5.72E-05	NaN
F ₅	NaN	6.39E-05	6.39E-05	2.17E-03	5.72E-05	NaN
F ₆	1.83E-04	2.20E-03	2.57E-02	1.83E-04	1.83E-04	5.83E-04
F ₇	1.41E-04	4.08E-01	5.02E-03	1.41E-04	6.04E-04	1.41E-04
F ₈	2.44E-02	1.68E-01	3.62E-01	2.44E-02	5.02E-03	5.31E-05

Table 6. Comparison of test results of different algorithms in Design of Reducer

Test algorithm	optimal value	The worst value	average value	standard deviation
DEDS	2994.4710	2994.4710	2994.4710	3.60E-12
ABC	2997.0584	N/A	2997.0584	0
NAFPSO	2994.4710	2994.4710	2994.4710	1.34E-12
DELC	2994.4710	2994.4710	2994.4710	1.90E-12
FWA	2904.1413	2980.3713	2925.2067	14.8455
MSFWA	2896.2593	2896.2593	2896.2593	3.01E-10

**Fig. 1.** Evolution curves for F3**Fig. 2.** Evolution curves for F4**Fig. 3.** ANOVA test for F6**Fig. 4.** ANOVA test for F8**Fig. 5.** Evolution curves for Design of Reducer**Fig. 6.** ANOVA test for Design of Reducer

References

1. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1942–1948. Perth Western Australia (1995)

2. Pan, W.: A new fruit fly optimization algorithm: taking the financial distress model as an example. *Knowl. Based Syst.* **26**, 69–74 (2012)
3. Tan, Y., Zhu, Y.C.: Fireworks algorithm for optimization. In: International Conference in Swarm Intelligence, pp: 355–364. Berlin (2010)
4. Zheng, S.Q., Janecek, A., Tan, Y.: Enhanced fireworks algorithm. In: IEEE Congress on Evolutionary Computation, pp: 2069–2077. Cancun, Mexico (2013)
5. Zheng, S.Q., Janecek, A., Li, J., et al.: Dynamic search in fireworks algorithm. In: IEEE Congress on Evolutionary Computation, Beijing, China. pp: 3222–3229 (2014)
6. Li, J., Zheng, S.Q., Tan, Y.: Adaptive Fireworks Algorithm. In: IEEE Congress on Evolutionary Computation, pp: 3214–3221. Beijing, China (2014)
7. Zhang, S.P., Wang, L.N.: Research and analysis on progress of fruit fly optimization algorithm. *Comput. Eng. Appl.* **57**(06), 22–29 (2021)
8. Zeng, M., Zhao, Z.G., Li, Z.M.: Self-learning improved fireworks algorithm with Cauchy mutation. *Journal of Chinese Computer Systems* **41**(02), 264–270 (2020)
9. Li, J., Zheng, S.Q., Tan, Y.: The effect of information utilization: Introducing a novel guiding spark in the fireworks algorithm. *IEEE Trans. Evol. Comput.* **14**(1), 153–166 (2017)
10. Li, J., Tan, Y.: Loser-out tournament-based fireworks algorithm for multimodal function optimization. *IEEE Trans. Evol. Comput.* **22**(5), 679–691 (2018)
11. Vapnik, V.N., Lerner, A.Y.: Recognition of patterns with help of generalized portraits. *Avtomat. i Telemekh.*, **24**(6), 774–780 (1963)
12. Arora, S., Singh, S.: Butterfly optimization algorithm: a novel approach for global optimization. *Soft. Comput.* **23**(3), 715–734 (2018). <https://doi.org/10.1007/s00500-018-3102-4>
13. Mirjalili, S., Lewis, A.: The Whale Optimization Algorithm. *Adv. Eng. Softw.* **95**, 51–67 (2016)
14. Gandomi, A.H., Yang, X.S., Alavi, A.H.: Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Eng. Comput.* **29**(1), 17–35 (2013)
15. Zhang, M., Luo, W.J., Wang, X.F.: Differential evolution with dynamic stochastic selection for constrained optimization. *Inf. Sci.* **178**(15), 3043–3074 (2008)
16. Akay, B., Karaboga, D.: Artificial bee colony algorithm for large-scale problems and engineering design optimization. *J. Intell. Manuf.* **23**(4), 1001–1014 (2012)
17. Su, S.B., Li, Z., He, C.: Constrained fractional-order PSO with self-adaptive neighbors and differential mutators. *J. Chongqing Univ.* **43**(11), 84–98 (2020)
18. Wang, L., Li, L.P.: An effective differential evolution with level comparison for constrained engineering design. *Struct. Multidiscip. Optim.* **41**(6), 947–963 (2010)



A New Fitness-Landscape-Driven Particle Swarm Optimization

Xuying Ji, Feng Zou^(✉), Debao Chen, and Yan Zhang

Huaibei Normal University, Huaibei 235000, China

zfemail@163.com

Abstract. Fitness landscape is an evolutionary mechanism and fitness landscape theory has developed considerably since it was proposed by Sewall Wright in the 1930s. In evolutionary algorithms, some characteristic information by analyzing the fitness landscape can be obtained to improve the optimization performance of algorithms. This paper introduces a new fitness-landscape-driven particle swarm optimization (FLDPSO). In the method, the correlation metric between fitness value and distance is obtained by characterizing the fitness landscape of optimization problems. Then, two new proposed variants of particle swarm optimization (PSO) are developed to improve the optimization performance. Moreover, a selection mechanism based on this metric is introduced to select a fitter variant from these two variants. Finally, the experimental simulation is executed on 18 benchmark functions to assess the optimization performance of the proposed FLDPSO algorithm. The results show that FLDPSO can improve optimization accuracy and convergence very well.

Keywords: Particle swarm optimization · Fitness landscape · Benchmark function

1 Introduction

Fitness landscape is an evolutionary mechanism proposed by Sewall Wright in the 1930s and applied to the field of biological evolution [1]. In biological evolution, if the corresponding relationship between genotype and fitness is described by “fitness terrain”, the fitness landscape can be regarded as a three-dimensional visualization landscape composed of mountains, plains and basins. In the past years, the fitness landscape has been utilized in evolutionary algorithms. In evolutionary algorithms, the fitness landscape is regarded as an evolutionary method which is randomly walking on a three-dimensional visual landscape to find the highest peak. In a word, the higher the peak, the greater the fitness value. With the application of fitness landscape in evolutionary algorithms, the feature information of fitness landscape such as local fitness of the landscape, fitness distance correlation and roughness of the landscape become more and more abundant. Furthermore, feature information of fitness landscape reflects the distribution of optimal solution, quantity of optimal solution and local unimodal extension structure of the optimization problem [2]. With the continuous attempts of researchers, fitness landscapes

have been introduced into evolutionary algorithms in various ways and used to solve various problems, such as introducing fitness landscapes into hysteretic systems to solve the problem of parameter estimation [3], introducing two fitness landscape analysis metrics into probabilistic matching operator selection scheme to deal with multi-objective optimization problems [4], and introducing the ruggedness of landscape to judge the topology of the local landscape and then combining the outcome with a reinforcement learning strategy to get the optimal probability distribution [5].

Particle swarm optimization (PSO) has been developed rapidly since it was first proposed by Eberhart and Kennedy [6]. The concept of the PSO algorithm comes from biological models such as ant colonies and birds. Each particle means a possible solution in the particle swarm, and all particles exchange their information to find the food which means the global optimum. Therefore, the algorithm has the characteristics of simplicity, few parameters to be adjusted, fast convergence speed, and easy implementation. However, it is also unstable and easy to sink into local optimum. Researchers have developed many variants to address these issues. Shi added an inertial weight parameter into the particle velocity and position updating equations to balance global and local search capabilities [7]. Clerc and Kennedy introduced a shrinkage factor that guaranteed and improved the convergence speed of PSO [8]. Peram et al. introduced a variant of PSO and this variant used the ratio of the fitness and the distance of other particles to guide the direction of particle position [9].

Section 2 describes the fitness distance correlation and the basic PSO algorithm. Section 3 introduces the new fitness-landscape-driven particle swarm optimization algorithm in detail, and the corresponding measurement results and comparison results are given in Sect. 4. Section 5 summarizes the improved FLDPSO algorithm and provide some directions in the future.

2 Related Works

The fitness distance correlation can reflect many properties of the optimization problem. This section gives the definition and formula of the fitness distance correlation and the basic particle swarm optimization is introduced.

2.1 Fitness Distance Correlation

There should be a certain correlation between individual fitness and distance. So, the conception of fitness distance correlation (FDC) method is the relationship between fitness of individuals and distance in the search space [2]. Hence, the properties of the optimization algorithm can be reflected by FDC. The formula of fitness distance correlation can be given as follow:

$$r_{FD} = \frac{c_{FD}}{s_D s_F} \quad (1)$$

$$c_{FD} = \frac{1}{S} \sum_{i=1}^S (f^{(i)} - \bar{f})(d^{(i)} - \bar{d}) \quad (2)$$

where S is the total number of individuals, S_D and S_F are the standard deviation of the distance and fitness values, respectively. \bar{f} represents the average fitness, \bar{d} is mean distance. The distance formula is as follows:

$$d(X_i, gbest) = \sum_{j=1}^n |X_{i,j} - gbest_j| \quad (3)$$

where n is the dimension, $gbest$ is the global optimum.

When the distance between the individual and the optimal individual is smaller, the corresponding fitness is also larger. This means that it is a unimodal problem. When it does not get larger with decreasing distance, it is a multimodal problem. In this means, the fitness distance correlation can reflect some characteristics of optimization problems. FDC was introduced into many algorithms to solve the optimization problem. Jones introduced FDC into the genetic algorithm (GA) to measure the difficulty of the problem [10]. Li et al. put FDC into differential evolution (DE) algorithm to select an appropriate strategy [2]. Collard et al. introduced FDC into the genetic algorithm (GA) to analyze the statistical measure [11].

2.2 Basic Particle Swarm Optimization

Particle swarm optimization (PSO) has developed rapidly since it was first proposed by Eberhart and Kennedy [6]. In fact, it is a process in which each particle exchanges information about the minimum value and jointly finds the global minimum value. The updating equations for the position x_i and velocity v_i of each particle are given as follows [12]:

$$\begin{cases} v_i(g+1) = v_i(g) + c_1 * rand * (pbest - x_i(g)) + c_2 * rand * (gbest - x_i(g)) \\ x_i(g+1) = x_i(g) + v_i(g+1) \end{cases} \quad (4)$$

where i represents the i -th particle in the population, g is the number of iterations, c_1 and c_2 are the constant parameters, $rand$ is a random number between $[0,1]$, $pbest_i$ is the historical optimal position of the i -th particle, $gbest$ is the optimal position of the entire population.

Each particle has its velocity and position, and they are guided and updated by the global and local optima. It can be seen from the formula (4) that when c_1 is large, the speed is very likely to approach $pbest$, and when c_2 is large, the speed is very likely to approach $gbest$.

3 Proposed FLDPSO Algorithm

The optimization problem can be better solved if its fitness landscape characteristic can be analyzed and then a much more appropriate optimization method is available adaptively. The correlation metric between fitness and distance can reflect the properties of the optimization problem. By utilizing these properties, a new fitness-landscape-driven particle swarm optimization (FLDPSO) is developed to promote the optimization performance of PSO. The proposed FLDPSO algorithm is described detailly in the following.

3.1 Fitness-Landscape-Driven Strategy

The conception of the correlation metric based on fitness distance is the relationship between fitness of individuals and distance in the search space. Hence, the properties of the optimization algorithm can be reflected by the relationship between fitness and distance of each individual in the search space. In the FLDPSO algorithm, the correlation metric (c_{FD}) based on fitness distance can be expressed as follow:

$$c_{FD} = (f^{(i)} - \bar{f})(d^{(i)} - \bar{d}) \quad (5)$$

where \bar{f} represents the average fitness, \bar{d} is the mean distance. The distance $d^{(i)}$ is given as follows:

$$d(X_i, gbest) = \sqrt{\sum_{j=1}^n (X_{i,j} - gbest_j)^2} \quad (6)$$

where n is the dimension, $gbest$ is the global optimum.

The difference between the fitness value of each individual and the average fitness value of all individuals can reflect the change direction of the fitness and the difference between each individual and the average individual reflect the change direction of the individual. The correlation metric c_{FD} reflects the properties of the entire optimization problem by calculating the change direction of each individual and its fitness. The change of the fitness and the distance are considered to have the same trend when they have the same direction. Hence, the correlation metric based on fitness distance can reflect the feature of the global fitness landscape of the optimization problem to a certain extent.

Based on the above analysis, in the FLDPSO algorithm, by counting the number of individuals with the same direction, a probability α can be determined so as to select an appropriate optimization method. Suppose that the number of particles in the population is S , the selection probability α can be obtained as follows.

Step1: Compute the average fitness value of each individual;

Step2: Compute the Euclidean distance between each individual and the current global optimal individual according to Eq. (6), and then obtain their average distance value;

Step3: Compute the correlation metric according to Eq. (5) for each individual;

Step3: Suppose $\theta = 0$. For each individual, $\theta = \theta + 1$ if $c_{FD} > 0$;

Step4: Normalize $\alpha = \theta/S$.

3.2 Variants of PSO

In the FLDPSO algorithm, two variants of PSO are designed to update individuals, and these two new updating formulas are considered as SD and SC as follow.

SD: In this method, the normally distributed individual, the neighbor optimal individual, and the average individual are introduced into PSO to update each individual as following formula:

$$\begin{cases} v_i(g+1) = v_i(g) + c_1 * rand * (pbest_i(g) - (x_i(g) + n_i(g))/2) \\ \quad + c_2 * rand * (pbest_c - (x_i(g) + n_i(g))/2) \\ \quad + rand * (gbest - TF * meanx_i(g)) \\ x_i(g+1) = x_i(g) + v_i(g+1) \end{cases} \quad (7)$$

where $pbest_c$ is individual optimum of its neighbors, $meanx_i(g)$ is the average individual in the population, TF is a heuristic step and can be either 1 or 2 randomly. $n_i(g)$ is the normally distributed individual as following formula [14]:

$$n_i(g) = normrnd((gbest(g) + pbest_i(g))/2, |gbest(g) - pbest_i(g)|) \quad (8)$$

SC: In this method, two updating methods are utilized according to the previous fitness value $gbest(g-1)$ and the current fitness value $gbest(g)$ of the global optimal individual.

Generally, if $gbest(g) = gbest(g-1)$, it means the population is stagnant and the population is updated as follow:

$$\begin{cases} v_i(g+1) = w * v_i(g) + c_1 * rand * (pbest - x_i(g)) + c_2 * rand * (gbest - x_i(g)) \\ x_i(g+1) = x_i(g) + v_i(g+1) \end{cases} \quad (9)$$

Otherwise, it means the global optima changes the better. Hence, the algorithm adopts PSO social model as follow:

$$\begin{cases} v_i(g+1) = v_i(g) + c_2 * rand * (gbest(g) - x_i(g)) \\ x_i(g+1) = x_i(g) + v_i(g+1) \end{cases} \quad (10)$$

For *SD*, three exemplar individuals are utilized to guide the searching of particle swarm to improve the diversity of the population, and thus helping to escape from the local optimum. For *SC*, the variability of the global optimum can reflect the evolution direction to a certain extent. Hence, the variability of the previous and current global optimum is judged to determine its updating direction and accelerate the convergence speed. Therefore, when the global optimum does not change, Eq. (9) is executed to update the particles. Otherwise, the PSO social model is adopted to accelerate the speed of the population towards the global optimum. That is, Eq. (10) is executed to update the particles.

3.3 Selection Strategy

The FLDPSO algorithm adopts a probability α to determine which operation is selected from SD and SC operations. Based on the above analysis, when less individuals have the same evolution trend and it reflect that the fitness landscape is closer to the multimodal local fitness landscape, SD is selected to improve the diversity of the population and escape from local optimum. When multiple individuals have the same properties and it reflect that the fitness landscape is closer to the unimodal local fitness landscape, SC is selected to increase the convergence and accuracy of the algorithm. Hence, the detail selection strategy based on the probability α can be given as follow:

```

if  $r < \alpha$ 
    Execute SC operction according to Eq. (9 – 10)
else
    Execute SD operction according to Eq. (7 – 8)

```

(11)

In a word, the possibility of choosing SC operation is greater when the value of α is larger, and the possibility of choosing SD operation is greater when the value of α is smaller.

4 Simulation

4.1 Evaluation Function and Parameter Setting

This section selects 18 benchmark functions [13] to test the performance of the algorithm, and compares and analyzes the performance of some classical PSO algorithms. FLDPSO is compared with 8 improved PSO algorithms, among which are BBPSO [14], CLPSO [15], PSOcf [8], PSOcfLocal [16], PSOFDR [9], PSOw [7], PSOwLocal [16]. All algorithms were performed under MATLABR2018b. The parameter settings of the algorithm are all the same as the original settings. The experiment is carried out on dimension 30, and the maximum evaluation value of each function is set to 150000, the population size is to set 50, and it runs 10 times independently.

4.2 Experimental Results

The test results of 30D on benchmark function test set are displayed in Table 1, including mean and variance. Here, we mark the best results of these algorithms on each test function in bold font.

From the table, the FLDPSO algorithm has obvious improvement over other algorithms on the F1, F2, F3, F4, F6, F8, F11, F12, F14, F15, F16 functions, and for the F1-F5 single-modal optimization problems and F11-F18 rotation optimization problems have very good performance. In the F6-F10 multimodal optimization problems, there has a little improvement. In general, the FLDPSO algorithm still has good performance

when dealing with optimization problems, and the proposed FLDPSO algorithm has more excellent convergence performance and better ability to escape from local optimum than other algorithms. This shows that the fitness-landscape-driven strategy can outstandingly guarantee that the FLDPSO algorithm converges to the global optimum efficiently.

Table 1. Computational results of benchmark functions with 30 dimensions

Name	Metric	F1	F2	F3	F4	F5	F6
BBPSO	mean	8.8487E-58	7.0013E + 03	1.5000E + 02	1.1593E + 01	2.3019E + 01	1.1551E-01
	Std	1.7853E-57	6.3231E + 03	1.8409E + 02	3.1392E + 01	1.8780E + 01	3.6529E-01
CLPSO	mean	1.6565E-24	1.0683E + 02	2.6782E-25	5.1080E-02	2.3889E + 01	3.1299E-13
	Std	2.1664E-24	5.6457E + 01	3.4076E-25	3.6392E-02	2.9611E + 00	1.7543E-13
PSOcf	mean	3.0194E-64	5.0000E + 02	1.0052E-64	1.0169E-10	1.7023E + 01	8.9477E-01
	Std	9.2707E-64	1.5811E + 03	2.3466E-64	1.7369E-10	2.1050E + 00	6.3552E-01
PSOcfLocal	mean	4.5335E-32	2.2539E-01	4.2038E-33	1.0522E-04	2.0734E + 01	7.1054E-15
	Std	3.3367E-32	1.1123E-01	5.5558E-33	8.6038E-05	2.3134E-01	0.0000E + 00
PSOFDR	mean	9.4278E-68	5.7237E-05	1.5094E-69	2.1813E-08	1.3405E + 01	7.1054E-15
	Std	2.9797E-67	5.5392E-05	4.7553E-69	3.7175E-08	1.5642E + 00	0.0000E + 00
PSOw	mean	6.2736E-19	1.3961E + 01	1.3724E-20	8.8928E-02	2.3846E + 01	1.8591E-10
	Std	1.0324E-18	5.4674E + 00	2.6543E-20	4.8173E-02	1.5179E-01	1.9633E-10
PSOwFIPS	mean	1.0598E-06	2.5029E + 02	1.1721E-07	8.5916E-01	2.5387E + 01	2.2937E-04
	Std	2.2100E-07	6.0029E + 01	3.3881E-08	2.7627E-01	1.7939E-01	4.3659E-05
PSOwLocal	mean	7.4985E-08	1.4898E + 02	4.3418E-09	1.5424E + 00	2.5324E + 01	6.2574E-05
	Std	1.1138E-07	3.7247E + 01	4.9837E-09	3.6572E-01	2.3995E-01	2.9885E-05
FLDPSO	mean	1.8742E-169	1.3853E-40	1.4300E-169	2.5499E-48	1.7848E + 01	3.5527E-15
	Std	0.0000E + 00	4.1726E-40	0.0000E + 00	5.9567E-48	1.0207E + 00	0.0000E + 00
Name	Metric	F7	F8	F9	F10	F11	F12
BBPSO	mean	1.0404E + 02	9.5034E-01	1.7916E-02	3.4660E + 03	9.5976E + 01	4.2897E + 01
	Std	3.0619E + 01	6.8528E-01	2.0089E-02	8.2110E + 02	2.3673E + 02	7.3794E + 01
CLPSO	mean	1.0447E + 01	1.8269E-05	2.3648E-15	2.5109E + 03	1.3792E-06	5.1081E-02
	Std	2.1621E + 00	4.0120E-05	6.1943E-15	3.7912E + 02	2.9251E-06	4.1783E-02
PSOcf	mean	5.9108E + 01	3.4007E + 00	8.6144E-03	4.9472E + 03	6.4107E-43	3.4953E-11
	Std	1.3153E + 01	2.8503E + 00	9.7221E-03	8.4732E + 02	1.9877E-42	5.0114E-11
PSOcfLocal	mean	3.2068E + 01	4.5187E-05	8.1278E-03	3.9548E + 03	1.0561E-14	5.6844E-05
	Std	6.3140E + 00	1.4290E-04	7.6110E-03	7.4623E + 02	2.8410E-14	3.9615E-05
PSOFDR	mean	2.6665E + 01	3.1126E-01	1.0589E-02	3.4568E + 03	1.6685E-43	4.9186E-09
	Std	4.6620E + 00	9.4589E-01	9.7100E-03	6.4594E + 02	5.2762E-43	7.1536E-09
PSOw	mean	2.9053E + 01	1.5017E-01	1.5258E-02	2.3708E + 03	2.5620E-09	9.9334E-02
	Std	4.7090E + 00	4.7428E-01	1.0480E-02	4.6578E + 02	8.0407E-09	4.7889E-02
PSOwFIPS	mean	4.9737E + 01	1.3245E-02	1.4957E-04	2.9303E + 03	9.9048E-06	8.2370E-01
	Std	9.0880E + 00	1.7858E-03	2.4960E-04	5.5048E + 02	1.1668E-05	3.1583E-01

(continued)

Table 1. (*continued*)

Name	Metric	F7	F8	F9	F10	F11	F12
PSOwLocal	mean	2.5971E + 01	1.4267E-03	2.9565E-03	4.3843E + 03	3.2741E-04	1.2300E + 00
	Std	4.2074E + 00	4.4317E-04	5.1664E-03	3.3929E + 02	8.8937E-04	6.4160E-01
FLDPSO	mean	1.5532E + 01	0.0000E + 00	1.3478E-13	3.7983E + 03	2.1517E-152	5.8782E-49
	Std	9.6587E + 00	0.0000E + 00	3.5935E-13	8.4315E + 02	6.8012E-152	9.8050E-49
Name	Metric	F13	F14	F15	F16	F17	F18
BBPSO	mean	8.9487E + 01	9.3979E-01	1.1860E + 02	1.2517E + 01	1.2784E-02	4.3830E + 03
	Std	4.6628E + 01	1.0094E + 00	3.6835E + 01	2.8771E + 00	1.1847E-02	6.9717E + 02
CLPSO	mean	3.9687E + 01	2.8848E-13	4.1203E + 01	3.0848E + 00	4.7381E-03	2.8227E + 03
	Std	1.6807E + 01	3.7035E-13	7.2668E + 00	2.4361E + 00	6.5394E-03	4.4393E + 02
PSOcf	mean	7.2364E + 01	1.5607E + 00	6.0818E + 01	6.1695E + 00	1.6237E-02	4.3672E + 03
	Std	5.6645E + 01	6.2821E-01	1.7566E + 01	1.8389E + 00	1.8982E-02	5.9534E + 02
PSOcfLocal	mean	2.3727E + 01	7.4607E-15	3.4824E + 01	1.9389E + 00	5.4204E-03	4.2089E + 03
	Std	2.8436E + 00	1.1235E-15	9.2149E + 00	1.2462E + 00	6.3374E-03	5.0713E + 02
PSOFDR	mean	3.1364E + 01	4.1586E-01	4.3281E + 01	1.6315E + 00	1.9188E-02	3.6122E + 03
	Std	8.3686E + 00	6.7622E-01	1.1529E + 01	1.0067E + 00	1.1682E-02	9.7335E + 02
PSOw	mean	4.4645E + 01	1.1334E + 00	5.2336E + 01	2.9246E + 00	1.6221E-02	2.8671E + 03
	Std	2.6026E + 01	7.7851E-01	1.3328E + 01	1.5572E + 00	1.5494E-02	4.4195E + 02
PSOwFIPS	mean	2.5530E + 01	2.9698E-04	1.1641E + 02	5.6631E-02	1.9276E-04	3.3413E + 03
	Std	9.5151E-01	7.5643E-05	1.3247E + 01	9.9963E-03	1.6660E-04	7.2649E + 02
PSOwLocal	mean	5.1512E + 01	2.2813E-01	2.6969E + 01	1.8704E + 00	1.7473E-03	3.8638E + 03
	Std	2.3630E + 01	4.8955E-01	5.0474E + 00	8.6704E-01	3.6907E-03	5.8384E + 02
FLDPSO	mean	4.4404E + 01	3.5527E-15	2.0198E + 01	0.0000E + 00	2.5704E-03	4.0903E + 03
	Std	2.2366E + 01	0.0000E + 00	1.1470E + 01	0.0000E + 00	5.6472E-03	4.6938E + 02

In order to better contrast the convergence of the FLDPSO algorithm and other algorithms, the convergence curve of each algorithm is shown in Fig. 1, and the FLDPSO algorithm is the red curve. It can be seen from the figure that for single-mode, multi-mode and rotation optimization problems, the convergence has been improved to a certain extent, especially for single-mode optimization problems, the convergence curve can be seen to be significantly improved. For multi-mode optimization problems, its convergence curve cannot be said to be the best compared to other algorithms. For rotational optimization problems, the part of the convergence curves is good.

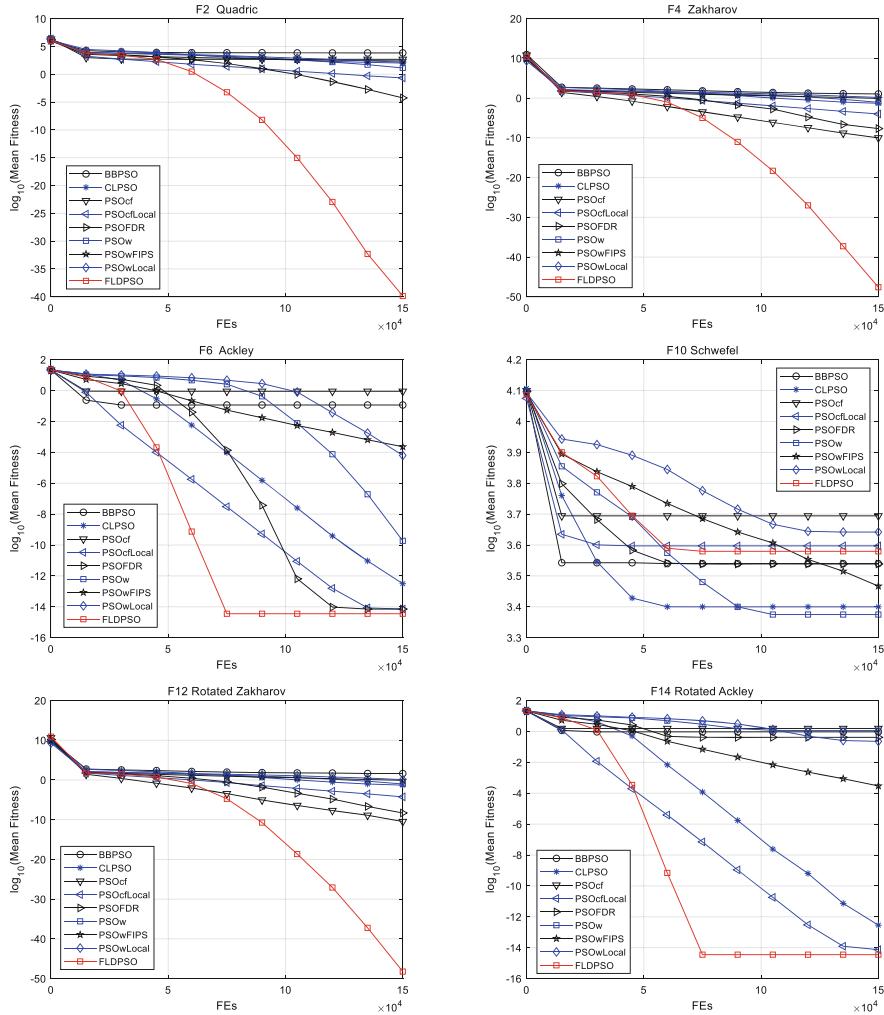


Fig. 1. Comparison of the performance curves of 9 algorithms on the 18 test functions for 30-D problems

5 Conclusions

This paper proposes a new FLDSO algorithm where it is based on the correlation between fitness and distance. And two new update formulas are used to solve the optimization algorithm. For the update formula, one is to guarantee the diversity of the population, and the other is to improve the speed of the convergence to the global optimum. The experimental results demonstrate that the FLDSO algorithm has good performance contrasted with other PSO algorithms, which ensures and increases the convergence speed and accuracy. For different properties of optimization problems, the strategy plays a good role in judging.

Although the difficulty of the problem is well analyzed through the fitness-landscape-driven strategy, and the algorithm solves the single-modal problems, multimodal optimization problems, and rotation optimization problems well in the experimental results, there are still some shortcomings that need to be improved. In future research, new representative methods of the fitness landscape feature need to be designed to better characterize the fitness-landscape information of problems. On the other hand, the fitness-landscape-driven particle swarm optimization would be utilized to deal with multimodal optimization problems, multi-objective optimization problems and dynamic optimization problems.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 61976101) and the funding plan for scientific research activities of academic and technical leaders and reserve candidates in Anhui Province (No. 2021H264).

References

1. Wright, S.: The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: Proceedings of the Sixth International Congress on Genetics, pp. 355–366 (1932)
2. Li, W., Meng, X., Huang, Y.: Fitness distance correlation and mixed search strategy for differential evolution. Neurocomputing **458**(2021), 514–525 (2020)
3. Worden, K., Manson, G.: On the identification of hysteretic systems. Part I: fitness landscapes and evolutionary identification. Mech. Syst. Signal Process. **29**, 201–212 (2012)
4. Kuk, J., Goncalves, R., Pozo, A.: Combining fitness landscape analysis and adaptive operator selection in multi and many-objective optimization. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), pp. 503–508. IEEE (2019)
5. Huang, Y., Li, W., Tian, F., et al.: A fitness landscape ruggedness multiobjective differential evolution algorithm with a reinforcement learning strategy. Appl. Soft Comput. **96**, 106693 (2020)
6. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: ICNN 1995-International Conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE (1995)
7. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence, pp. 69–73. IEEE (1998)
8. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evol. Comput. **6**(1), 58–73 (2002)
9. Peram, T., Veeramachaneni, K., Mohan, C.K.: Fitness-distance-ratio based particle swarm optimization. In: Proceedings of the 2003 IEEE Swarm Intelligence Symposium, pp. 174–181. IEEE (2003)
10. Jones, T.C., Forrest, S.: Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Proceedings of the Sixth International Conference on Genetic Algorithms, vol. 95, pp. 184–192 (1995)
11. Collard, P., Gaspar, A., Clergue, M., et al.: Fitness distance correlation, as statistical measure of genetic algorithm difficulty, revisited. In: (ECAI) European Conference on Genetic Programming, pp. 650–654 (1998)
12. Kameyama, K.: Particle swarm optimization-a survey. IEICE Trans. Inf. Syst. **92**(7), 1354–1361 (2009)
13. Zou, F., Wang, L., Hei, X., et al.: Bare bones teaching-learning-based optimization. Sci. World J. Article ID 136920, 17 (2014)

14. Kennedy, J.: Bare bones particle swarms. In: Proceedings of the 2003 IEEE Swarm Intelligence Symposium, pp. 80–87. IEEE (2003)
15. Liang, J.J., Qin, A.K., Suganthan, P.N., et al.: Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. *IEEE Trans. Evol. Comput.* **10**(3), 281–295 (2006)
16. Kennedy, J., Mendes, R.: Population structure and particle swarm performance. In: Proceedings of the 2002 Congress on Evolutionary Computation, vol. 2, pp. 1671–1676 . IEEE (2002)



Neighborhood Combination Strategies for Solving the Bi-objective Max-Bisection Problem

Rong-Qiang Zeng^{1,2(✉)} and Matthieu Basseur³

¹ Department of Computer Science, Chengdu University of Information Technology,
Chengdu 610225, Sichuan, People's Republic of China
zrq@cuit.edu.cn

² Chengdu Documentation and Information Center, Chinese Academy of Sciences,
Chengdu 610041, Sichuan, People's Republic of China

³ LERIA, Université d'Angers, 2, Boulevard Lavoisier, 49045 Angers Cedex 01, France
basseur@info.univ-angers.fr

Abstract. Local search is known to be a highly effective metaheuristic framework for solving a large number of classical combinatorial optimization problems, which strongly depends on the characteristics of neighborhood structure. In this paper, we integrate the neighborhood combination strategies into the hypervolume-based multi-objective local search algorithm, in order to solve the bi-criteria max-bisection problem. The experimental results indicate that certain combinations of neighborhood strategies are superior to others and the performance analysis sheds lights on the ways to further improvements.

Keywords: Multi-objective optimization · Hypervolume contribution · Local search · Neighborhood combination · Max-bisection problem

1 Introduction

Given an undirected graph $G = (V, E)$ with the vertex set $V = \{1, \dots, n\}$ and the edge set $E \subset V \times V$. Each edge $(i, j) \in E$ is associated with a weight w_{ij} . The max-cut problem is to seek a partition of the vertex set V into two disjoint subsets V_1 and V_2 , which is mathematically formulated as follows [1].

$$f_k(V_1, V_2) = \max \sum_{i \in V_1, j \in V_2} w_{ij}^k \quad (1)$$

where w_{ij}^k is the weight of the k^{th} ($k \in \{1, 2\}$) graph.

When the two subsets V_1 and V_2 are required to have the same cardinality (assuming that n is even), the bi-criteria max-cut problem becomes the bi-criteria max-bisection problem [9].

As the max-cut problem is one of Karp's 21 NP-complete problems [4], the max-bisection problem remains NP-complete in the general case [9]. a number of heuristics

and metaheuristics have been proposed to tackle this problem, including tabu search [6], advanced scatter search [7], global equilibrium search [8], etc.

Generally, local search is a simple and effective metaheuristic framework for solving a large number of classical combinatorial optimization problems, which proceeds from an initial solution with a sequence of local changes by defining the proper neighborhood structure for the considered problem. In order to study different neighborhood combination strategies during the local search process, we present the experimental analysis of the neighborhoods to deal with the bi-criteria max-bisection problem.

In this paper, we integrate the neighborhood combination strategies into the hypervolume-based multi-objective local search algorithm, in order to study the search capability of different neighborhood combinations on the bi-criteria max-bisection problem. The experimental results indicate that certain combinations of neighborhood strategies are superior to others. The performance analysis explains the behavior of the algorithms and sheds lights on the ways to further enhance the search process.

The remaining part of this paper is organized as follows. In the next section, we briefly introduce the basic notations and definitions of bi-objective optimization. Then, we present the hypervolume-based multi-objective local search algorithm with the neighborhood combination strategies for solving bi-criteria max-bisection problem in Sect. 3. Section 4 indicates that the experimental results on the benchmark instances of max-bisection problem. The conclusions are provided in the last section.

2 Bi-objective Optimization

In this section, we briefly introduce the basic notations and definitions of bi-objective optimization.

Without loss of generality, we assume that X denotes the search space of the optimization problem under consideration and $Z = \Re^2$ denotes the corresponding objective space with a maximizing vector function $Z = f(X)$, which defines the evaluation of a solution $x \in X$ [5]. Specifically, the dominance relations between two solutions x_1 and x_2 are presented below [11].

Definition 1. (Pareto Dominance). A decision vector x_1 is said to dominate another decision vector x_2 (written as $x_1 \succ x_2$), if $f_i(x_1) \geq f_i(x_2)$ for all $i \in \{1, 2\}$ and $f_j(x_1) > f_j(x_2)$ for at least one $j \in \{1, 2\}$.

Definition 2. (Pareto Optimal Solution). $x \in X$ is said to be Pareto optimal if and only if there does not exist another solution $x' \in X$ such that $x' \succ x$.

Definition 3. (Non-Dominated Solution). $x \in S$ ($S \subset X$) is said to be non-dominated if and only if there does not exist another solution $x' \in S$ such that $x' \succ x$.

Definition 4. (Pareto Optimal Set). S is said to be a Pareto optimal set if and only if S is composed of all the Pareto optimal solutions.

Definition 5. (Non-Dominated Set). S is said to be a non-dominated set if and only if any two solutions $x_1 \in S$ and $x_2 \in S$ such that $x_1 \not\succ x_2$ and $x_2 \not\succ x_1$.

Actually, we are interested in finding the Pareto optimal set, which keeps the best compromise among all the objectives. However, it is very difficult or even impossible to generate the Pareto optimal set in a reasonable time for the NP-hard problems. Therefore, we aim to obtain a non-dominated set which is as close to the Pareto optimal set as possible. That's to say, the whole goal is to identify a Pareto approximation set with high quality.

3 Neighborhood Combination Strategies

In this work, we integrate the neighborhood combination strategies into the hypervolume-based multi-objective local search algorithm, in order to deal with the bi-criteria max-bisection problem.

The general scheme of Hypervolume-Based Multi-Objective Local Search (HBMOLS) algorithm [3] is presented in Algorithm 1, and the main steps of this algorithm are described in the following subsections.

Algorithm 1 Hypervolume-Based Multi-Objective Local Search Algorithm

```

01: Input:  $N$  (Population size)
02: Output:  $A$ : (Pareto approximation set)
03: Step 1 - Initialization:  $P \leftarrow N$  randomly generated individuals
04: Step 2:  $A \leftarrow \emptyset$ 
05: Step 3 - Fitness Assignment: Assign a fitness value to each individual  $x \in P$ 
06: Step 4:
07: While Running time is not reached do
08:   repeat:
09:     Hypervolume-Based Local Search:  $x \in P$ 
10:   until all neighbors of  $x \in P$  are explored
11:    $A \leftarrow$  Non-dominated individuals of  $A \cup P$ 
12: end While
13: Step 5: Return  $A$ 
```

In HBMOLS, each individual in an initial population is generated by randomly assigning the vertices of the graph to two vertex subsets V_1 and V_2 with equal number. Then, we use the Hypervolume Contribution (HC) indicator defined in [3] to realize the fitness assignment for each individual.

Based on the dominance relation, the HC indicator divides the whole population into two sets: non-dominated set and dominated set, and calculates the hypervolume contribution of each individual in the objective space, according to two objective function values.

Actually, the individual belonging to the non-dominated set is assigned a positive value, while the individual belonging to dominated set is assigned to a negative value. Afterwards, each individual is optimized by the hypervolume-based local search procedure, which is presented in Algorithm 2 below.

Algorithm 2 Hypervolume-Based Local Search**Steps:**

- 01: $x^* \leftarrow$ an unexplored neighbor of x by randomly changing the values of the variables of x
- 02: $P \leftarrow P \cup x^*$
- 03: calculate two objective function values of x^*
- 04: calculate the fitness value of x^* in $\$P\$$ with the HC indicator
- 05: update all the fitness values of $z \in P (z \neq x^*)$
- 06: $\omega \leftarrow$ the worst individual in P
- 07: $P \leftarrow P \setminus \{\omega\}$
- 08: update all the fitness values of $z \in P$
- 09: if $\omega \neq x^*$, Progress \leftarrow True

In the hypervolume-based local search procedure, we implement the f -flip ($f \in \{1, 2\}$) move based neighborhood strategy and the combination. More Specifically, we randomly select two vertices from two sets and put them into the opposite set, so as to obtain an unexplored neighbor x^* of the individual x .

According to the HC indicator, a fitness value is assigned to this new individual. the individual ω with the worst fitness value will be deleted from the population P . The whole population is optimized by the hypervolume-based local search procedure, which will repeat until the termination criterion is satisfied, in order to generate a high-quality Pareto approximation set.

3.1 One-Flip Move

In order to generate a neighbor individual of the max-bisection problem, one-flip move is employed to achieve this goal by moving two randomly selected vertices to the opposite set, which is calculated as follows:

$$\Delta_i = \sum_{x \in V_1, x \neq v_i} w_{v_i x} - \sum_{y \in V_2} w_{v_i y}, v_i \in V_1 \quad (2)$$

$$\Delta_i = \sum_{x \in V_2, x \neq v_i} w_{v_i x} - \sum_{y \in V_1} w_{v_i y}, v_i \in V_2 \quad (3)$$

Let Δ_i be the move gain of representing the change of one vertex from V_1 to V_2 (or from V_1 to V_2) in the fitness function, and Δ_i can be calculated in linear time by the formula above, more details about this formula can be found in [10]. Then, we can calculate the objective function values high efficiently with the streamlined incremental technique.

3.2 Two-Flip Move

In the case of two-flip move, we can obtain a new neighbor individual by randomly moving four different vertices $v_a \in V_1$, $v_b \in V_1$ and $v_c \in V_2$, $v_d \in V_2$ to the opposite set. In fact, two-flip move can be seen as a combination of two single one-flip moves.

We denote the move value by δ_{ij} , which is derived from two one-flip moves Δ_i and Δ_j ($i \neq j$) as follows:

$$\delta_{ij} = \Delta_i + \Delta_j \quad (4)$$

Especially, the search space generated by two-flip move is much bigger than the one generated by one-flip move. In the following, we denote the neighborhoods with one-flip move and two-flip move as N_1 and N_2 respectively.

4 Experiments

In this section, we present the experimental results of 3 neighborhood combination strategies on 9 groups of benchmark instances of max-bisection problem. All the algorithms are programmed in C++ and compiled using Dev-C++ 5.0 compiler on a PC running Windows 10 with Core 2.50 GHz CPU and 4 GB RAM.

4.1 Parameters Settings

In order to carry out the experiments on the bi-criteria max-bisection problem, we use two single-objective benchmark instances of max-cut problem with the same dimension provided in [9]¹ to generate one bi-objective max-bisection problem instance. All the instances used for experiments are presented in Table 1 below.

Table 1. Single-objective benchmark instances of max-cut problem used for generating bi-objective max-bisection problem instances.

	Dimension	Instance 1	Instance 2
bo_mbp_800_01	800	g1.rud	g2.rud
bo_mbp_800_02	800	g11.rud	g12.rud
bo_mbp_800_03	800	g15.rud	g19.rud
bo_mbp_800_04	800	g17.rud	g21.rud
bo_mbp_2000_01	2000	g22.rud	g23.rud
bo_mbp_2000_02	2000	g32.rud	g33.rud
bo_mbp_2000_03	2000	g35.rud	g39.rud
bo_mbp_1000_01	1000	g43.rud	g44.rud
bo_mbp_3000_01	3000	g49.rud	g50.rud

In addition, the algorithms need to set a few parameters, we only discuss two important ones: the running time and the population size, more details about the parameter settings for multi-objective optimization algorithms can be found in [2, 9]. The more information about the parameter settings in our algorithms is presented in the following Table 2.

¹ More information about the benchmark instances of max-cut problem can be found on this website: [https://www.stanford.edu/\\$sim\\$yyye/yyye/Gset/](https://www.stanford.edu/simyyye/yyye/Gset/).

Table 2. Parameter settings used fzs bi-criteria max-bisection problem instances: instance dimension (D), vertices (V), edges (E), population size (P) and running time (T).

	Dimension (D)	Vertices (V)	Edges (E)	Population (P)	Time (T)
bo_mbp_800_01	800	800	19176	20	40 [“]
bo_mbp_800_02	800	800	1600	20	40 [“]
bo_mbp_800_03	800	800	4661	20	40 [“]
bo_mbp_800_04	800	800	4667	20	40 [“]
bo_mbp_2000_01	2000	2000	19990	50	100 [“]
bo_mbp_2000_02	2000	2000	4000	50	100 [“]
bo_mbp_2000_03	2000	2000	11778	50	100 [“]
bo_mbp_1000_01	1000	1000	9990	25	50 [“]
bo_mbp_3000_01	3000	3000	6000	75	150 [“]

4.2 Performance Assessment Protocol

In this paper, we evaluate the efficacy of 3 different neighborhood combination strategies with the performance assessment package provided by Zitzler et al². The quality assessment protocol works as follows: First, we create a set of 20 runs with different initial populations for each strategy and each benchmark instance of max-cut problem. Then, we generate the reference set RS^* based on the 60 different sets A_0, \dots, A_{59} of non-dominated solutions.

According to two objective function values, we define a reference point $z = [r_1, r_2]$, where r_1 and r_2 represent the worst values for each objective function in the reference set RS^* . Afterwards, we assign a fitness value to each non-dominated set A_i by calculating the hypervolume difference between A_i and RS^* . Actually, this hypervolume difference between these two sets should be as close to zero as possible [12], an example is illustrated in Fig. 1.

4.3 Computational Results

In this subsection, we present the computational results on 9 groups of bi-criteria max-bisection problem instances, which are obtained by three different neighborhood combination strategies. The information about these algorithms are described in the table below:

In Table 3, the algorithms HBMOLS_($N_1 \cup N_2$) selects one of the two neighborhoods to be implemented at each iteration during the local search process, choosing the neighborhood N_1 with a predefined probability p and choosing N_2 with the probability $1 - p$. In our experiments, we set the probability $p = 0.5$.

² More information about the performance assessment package can be found on this website: <http://www.tik.ee.ethz.ch/pisa/assessment.html>.

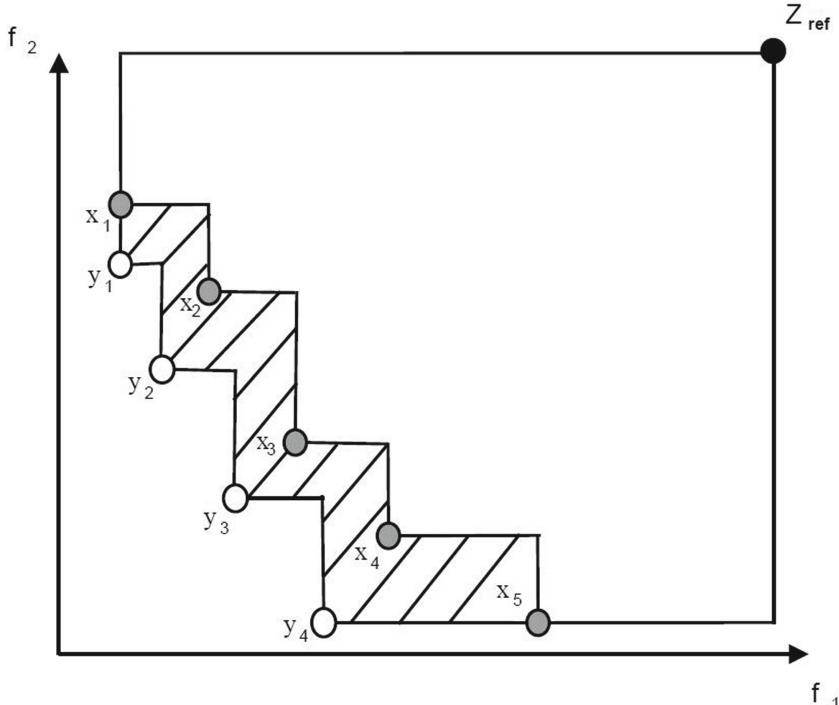


Fig. 1. Illustration of hypervolume difference (the shaded area) between a non-dominated set A_i (with 5 solutions $x_i, i \in \{1, \dots, 5\}$) and a reference set RS^* (with 4 solutions $y_i, i \in \{1, \dots, 4\}$).

Table 3. The algorithms with the neighborhood combination strategies.

	Algorithm description
HBMOLS_ N_1	one-flip move based local search
HBMOLS_ N_2	two-flip move based local search
HBMOLS_ $(N_1 \cup N_2)$	f -flip move based local search ($f \in \{1, 2\}$)

The computational results are summarized in Table 4. In this table, there is a value both **in bold** and **in grey box** at each line, which is the best result obtained on the considered instance. The values both **in italic** and **bold** at each line refer to the corresponding algorithms which are **not** statistically outperformed by the algorithm obtaining the best result (with a confidence level greater than 95%).

Table 4. The computational results on bi-criteria max-bisection problem obtained by the algorithms with 3 different neighborhood combination strategies.

Instances	Algorithms		
	N_2	$N_1 \cup N_2$	N_1
bo_mbp_800_01	0.144389	0.132989	0.128909
bo_mbp_800_02	0.140560	0.138421	0.133499
bo_mbp_800_03	0.116481	0.115812	0.101597
bo_mbp_800_04	0.110223	0.109617	0.091346
bo_mbp_2000_01	0.604618	0.589033	0.565971
bo_mbp_2000_02	0.620013	0.584992	0.579419
bo_mbp_2000_03	0.684170	0.653463	0.568310
bo_mbp_1000_01	0.154316	0.139311	0.111531
bo_mbp_3000_01	0.119639	0.116802	0.091568

From Table 4, we can observe that all the best results are obtained by N_1 , which statistically outperforms the other two algorithms on all the instances except for two instances (bo_mbp_2000_01 and bo_mbp_2000_02). Moreover, the results obtain by $N_1 \cup N_2$ is close to the results obtained by N_1 . Especially, the most significant result is achieved on the instance bo_mbp_2000_03, where the average hypervolume difference value obtained by N_1 is much smaller than the values obtained by the other two algorithms.

However, N_2 does not perform as well as N_1 , although the search space of N_2 is much bigger than N_1 . We suppose that there exists a number of key vertices in the representation of the individuals, which means these vertices should be fixed in one of two sets in order to search the local optima effectively. Two-flip move much more frequently changes the positions of these key vertices during the search process, so that the efficiency of local search of N_2 is obviously affected by the neighborhood strategy.

On the other hand, $N_1 \cup N_2$ provides another possibility to keep the positions of these key vertices unchanged, and broaden the search space at the same time. Then, the combination of one-flip move and two-flip move is very potential to obtain better results.

5 Conclusions

In this paper, we have presented the neighborhood combination strategies to solve the bi-criteria max-bisection problem, which are based on one-flip, two-flip and the combination. To achieve this goal, we have carried out the experiments on 9 groups of benchmark instances of max-bisection problem. The experimental results indicate that the better outcomes are obtained by the simple one-flip move based neighborhood and the neighborhood combination with two-flip is very potential to escape the local optima for further improvements.

Acknowledgments. The work in this paper was supported by the Fundamental Research Funds for the Central Universities (Grant No. A0920502051728–53) and supported by the West Light Foundation of Chinese Academy of Science (Grant No: Y4C0011006).

References

1. Angel, E., Gourves, E.: Approximation algorithms for the bi-criteria weighted max-cut problem. *Discret. Appl. Math.* **154**, 1685–1692 (2006)
2. Basseur, M., Lefooghe, A., Le, K., Burke, E.: The efficiency of indicator-based local search for multi-objective combinatorial optimisation problems. *J. Heuristics* **18**(2), 263–296 (2012)
3. Basseur, M., Zeng, R.-Q., Hao, J.-K.: Hypervolume-based multi-objective local search. *Neural Comput. Appl.* **21**(8), 1917–1929 (2012)
4. Benlic, U., Hao, J.-K.: Breakout local search for the max-cut problem. *Eng. Appl. Artif. Intell.* **26**, 1162–1173 (2013)
5. Coello, C.A., Lamont, G.B., Van Veldhuizen, D.A.: Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2007). <https://doi.org/10.1007/978-0-387-36797-2>
6. Kochenberger, G.A., Glover, F., Hao, J.-K., Lü, Z., Wang, H., Glover, F.: Solving large scale max cut problems via tabu search. *J. Heuristics* **19**, 565–571 (2013)
7. Martí, R., Duarte, A., Laguna, M.: Advanced scatter search for the max-cut problem. *Informs J. Comput.* **21**(1), 26–38 (2009)
8. Shylo, V.P., Shylo, O.V.: Solving the maxcut problem by the global equilibrium search. *Cybern. Syst. Anal.* **46**(5), 744–754 (2010)
9. Wu, Q., Hao, J.-K.: Memetic search for the max-bisection problem. *Comput. Oper. Res.* **40**, 166–179 (2013)
10. Wu, Q., Wang, Y., Lü, Z.: A tabu search based hybrid evolutionary algorithm for the max-cut problem. *Appl. Soft Comput.* **34**, 827–837 (2015)
11. Zitzler, E., Künzli, S.: Indicator-based selection in multiobjective search. In: Yao, X., et al. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 832–842. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30217-9_84
12. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *Evol. Comput.* **3**, 257–271 (1999)

Neural Networks



Rolling Bearing Fault Diagnosis Based on Model Migration

Yuchen Xing¹ and Hui Li^{1,2(✉)}

¹ School of Mechanical Engineering, Tianjin University of Technology and Education,
Tianjin 300222, China
Huili68@163.com

² Tianjin Key Laboratory of Intelligent Robot Technology and Application, Tianjin 300222,
China

Abstract. A rolling bearing fault diagnosis method based on deep transfer learning was proposed to solve the problems of low efficiency of rolling bearing fault classification under variable working conditions, complex model and traditional machine learning that could not adapt to weak calculation and less label. Firstly, the preprocessed data is used as the input layer of the one-dimensional convolutional neural network, and the learning rate multi-step attenuation strategy is used to train the model and construct the optimal model. Secondly, the optimal model is used to complete the rolling bearing fault classification in the target domain. Finally, compared with the ResNet model and TCA algorithm, the experimental results show that the proposed method has higher fault diagnosis accuracy than the ResNet model and TCA method, and is an effective method for automatic fault feature extraction and classification recognition.

Keywords: Model migration · Deep transfer learning · Fault diagnosis · Rolling bearing · Signal processing

1 Introduction

In the field of modern industry, the operation of the rotating parts directly affect the quality of industrial products and the production efficiency, among them, the rolling bearing, which is the core of the rotating machinery parts and components. In the event of failure will affect the operation of machinery and equipment. Therefore, research of the fault diagnosis of rolling bearing to keep the security and stability of machinery and equipment has the vital significance [1–3]. Traditional machine learning methods have been widely used in this field. However, with the increase of training data, when the available annotation data is reduced, its generalization ability in practical industrial applications is weak. In order to effectively solve the problem, low efficiency of traditional machine learning training data as well as increasingly lower ability to adapt to new data changes, scholars proposed a fault diagnosis method based on transfer learning (TL), which can find out labeled data from correlative domains for training when the data with label in the target domain is small. TL does not require the assumption

of the same distribution of training data and test data, which can effectively make the manpower and material resources reduced. It consumed by traditional machine learning methods to re-calibrate the obtained data [4]. Therefore, TL has been intensively used in the field of fault diagnosis. Shen et al. [5] show the method of feature extraction, which combined autocorrelation matrix SVD with TL to achieve motor fault diagnosis. Chen et al. [6] show a TL method of least squares Support Vector Machine (SVM) as well as the proposed model improved bearing diagnosis performance, but it was difficult to effectively implement for large-scale training samples. Von et al. [7] proposed the stationary subspace analysis (SSA) to achieve spatial matching of data with different distributions. Shi et al. [8] used information theoretical learning (ITL) to measure the similarity between different data samples. Gong et al. [9] used the geodesic flow kernel (GFK) method to measure the geometric distance of different distributed data in the Grassmann manifold space. Although the research results which have been mentioned have meetsatisfactory requirements, the two distance measurement methods still have some defects, such as large computation, high complexity and low generality. The problem, in the variable working condition of rolling bearing fault diagnosis as well as bearing fault vibration signals do not meet the assumption of the same distribution and it is difficult to mark the newly obtained working condition data. The main idea of TL is to learn from the source domain, then transfer the knowledge to the target domain so that complete the classification of the target domain. Transfer component analysis (TCA) maps different feature samples using kernel function to obtain the dimensionality reduction data of source domain and target domain. Moreover, this method which can effectively improve the data clustering and distinguishable between classes has the property of local geometric structure preservation.

In this paper, it proposes a bearing fault diagnosis method based on model for the low efficiency and complex model of fault diagnosis classification of mechanical equipment rotating parts under working conditions, which can automatically extract the characteristics of data samples to ensure the accuracy of classification. It can meet the practical application requirements of end to end. It can reduce the prior knowledge of data samples and further improve the adaptability of model-based migration in the field of fault diagnosis. Firstly, the traditional signal processing method is used to extract the input signal according to the fixed signal length and divide the source domain and target domain. CNN was used to extract the input signal characteristics and monitor the training process. Then, forward propagation and back propagation were used to modify the CNN model to complete the training process. Finally, the classifier is used to classify the target domain to verify the accuracy and validity of the optimal model.

2 Introduction of Transfer Learning

2.1 Transfer Learning

In the area of TL, the domain where labels are available is termed as source domain D_s , and the domain in which labels are not available is called target domain D_t , which is defined as follows:

$$D_t = \{(x_i^t)\}_{i=1}^{n_t} \quad x_i^t \in X_t \quad (1)$$

where D_t is the target domain, $x_i^t \in X_t$ is the sample i of the target domain, X_t is the union of all samples, and n_t is the total number of target samples.

2.2 Methods of Transfer Learning

The model-based migration method assumes that the samples of source domain and target domain can share model parameters, and the shared parameter information can be found from the source domain and target domain to realize the method of migration.

Loss Function. The source domain data are used to train CNN, wherein the loss of CNN classification adopts cross entropy loss, namely:

$$L = -\frac{1}{m} \sum_{i=1}^m p(x_{ij}) \log(q(x_{ij})) \quad (2)$$

where, m represents the number of samples, $p(x_{ij})$ represents the real distribution of samples and $q(x_{ij})$ represents the distribution predicted by the model.

A rolling bearing fault diagnosis schematic diagram based on model migration is shown by Fig. 1. Firstly, source domain and target domain are divided. Secondly, the pre-processed data are sent to the neural network for iteration and training to complete forward propagation.

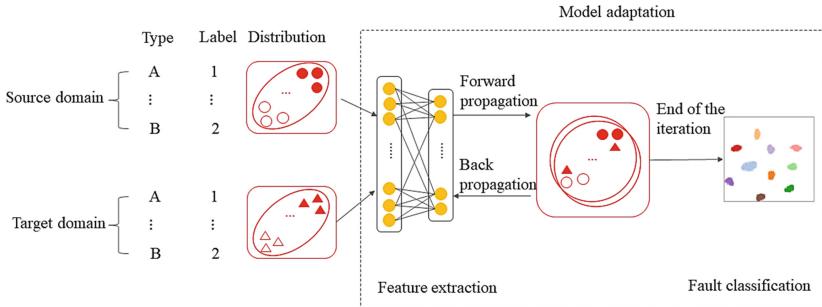


Fig. 1. Principle diagram of rolling bearing fault diagnosis based on model migration

Forward Propagation. Feature extraction of sample data is carried out through the convolution layer. Set as the output of 1 convolution layer, then 1D convolution can be expressed as:

$$x_j^l = f \left(b_j^l + \sum_i x_i^{l-1} * w_{i,j}^l \right) \quad (3)$$

where, x_i^{l-1} is the output of the neuron in number i at the layer in rank $l - 1$, b_j^l is the bias of the neuron in number j at the layer in rank l , $w_{i,l}^l$ is the convolution kernel from the neuron in number i at the layer in rank $l - 1$ to the neuron in number j at the 1th layer, $*$ is the convolution operator, $f(\cdot)$ is the activation function. ReLU has many advantages, fast convergence, small computation and no gradient loss, so that can be adopted in this paper.

Pooling layer is used to compress the input features, which can make the features clearer and simplify the computation of neural network on the one hand. Let the pooling size be $n * n$, and the layer $l - 1$ can be obtained by pooling the layer l , which can be obtained by the following formula:

$$x_i^l = f\left(\beta_j^l \text{down}\left(x_j^{l-1}\right) + b_j^l\right) \quad (4)$$

where, $\text{down}(\ast)$ is the lower sampling function, β_j^l represents the weight of the j feature graph at the l layer, and b_j^l represents the bias of the j feature graph at the l layer. After down-sampling, the number of feature maps owned by l down-sampling layer and $l - 1$ convolution layer remains unchanged, but the size is reduced $\frac{1}{n}$ to the original.

Back Propagation. The gradient descent method is adopted to adjust the network adaptive parameters, as shown in the following formula:

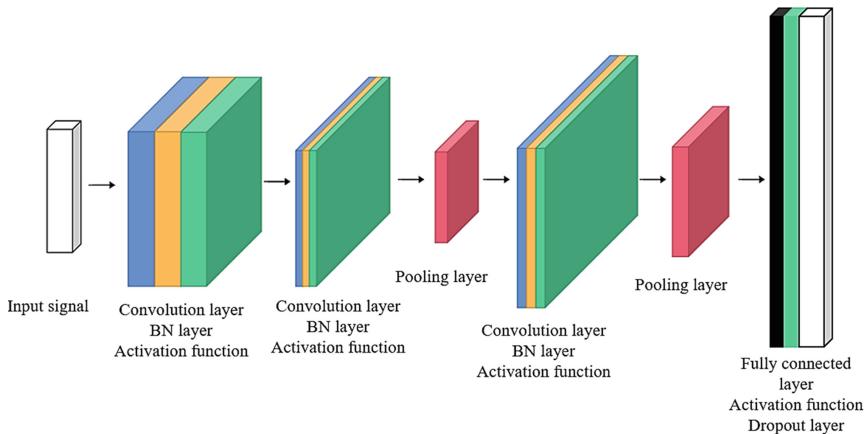
$$\left. \begin{array}{l} W_{pq}^{(l)}' = W_{pq}^{(l)} - \eta \frac{\delta}{\delta W_{pq}^{(l)}} L(D_s, D_t) \\ b_p^{(l)}' = b_p^{(l)} - \eta \frac{\delta}{\delta b_p^{(l)}} L(D_s, D_t) \end{array} \right\} \quad (5)$$

where, $W_{ij}, b_p^{(l)'}$ is the weight, p and q respectively represent the p node of the l layer and the q node of the $l + 1$ layer, and η is the learning rate.

After iteration, the trained model is obtained and saved for fault classification in the target domain.

2.3 1D CNN Model and Parameters

In this paper, the CNN model adopted is composed of input layer, convolution layer, BN layer, fully connected layer, Dropout layer, and pooling layer, as shown in Fig. 2. The network structure parameters are shown in Table 1. The network structure parameters about 1D CNN are shown in Table 1. 1D CNN adopts a lightweight network structure. The whole network only includes one fully connected layer, 1 dropout layer, two pooling layers and four convolution layers. This lightweight network structure not only simplifies the network structure, reduces the network parameters and increases the calculation speed, but also effectively avoids the phenomenon of over fitting.

**Fig. 2.** Structure of 1D CNN**Table 1.** Network structure parameters

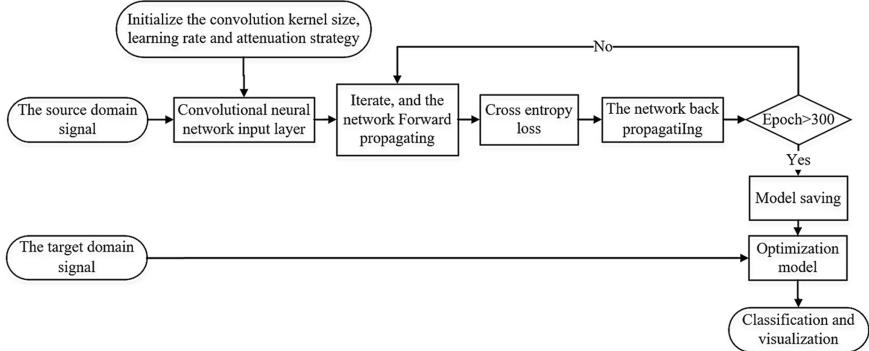
Number	Name	Size of convolution kernel	Number of convolution kernel
1	Convolution layer	15	16
2	Convolution layer	3	32
3	Pooling layer	2	32
4	Convolution layer	3	64
5	Convolution layer	3	128
6	Adaptive pooling layer	/	/
7	Fully connected layer	/	/
8	Dropout layer	Dropout rate = 0.5	

2.4 Fault Diagnosis Procedure Based on Model Migration

The process in fault diagnosis of bearing based on model migration is shown in Fig. 3.

The major steps are given as follows:

- (1) Obtaining the original vibration signal, and divide the input signal into source domain training set and target domain verification set.
- (2) Setting up diagnostic parameters, such as iteration times, initial learning rate, etc.
- (3) Sending the input data to the diagnostic model in batch for deep migration training.
- (4) Repeating step (3) until all the source domain data sets are trained and the optimal model is saved.
- (5) Inputting the data set in target domain and outputting the diagnosis results.
- (6) Visualizing bearing fault mode classification.

**Fig. 3.** 1D CNN based flow chart of bearing fault diagnosis

3 Bearing Fault Diagnosis Based on Model Migration

3.1 Experimental Data

In order to verify the accuracy and effectiveness of model transfer-based rolling bearing fault diagnosis, case western reserve university (CWRU) [11] rolling bearing data set was used in this paper. The sampling frequency of the experiment is 12 kHz, and the sampling target is the driving end rolling bearing, model SKF6250. The experimental platform bearing speed is 1797 rpm and 1730 rpm, which are the original vibration signals of the source domain and target domain of transfer learning. The experimental bearing fault defect is caused by EDM, and the fault degree is mild damage respectively. The damage diameter was 7 mils, 14 mils, and 21 mils(1 mils = 0.001 inches). The damage locations were the rolling body, bearing inner ring, and bearing outer ring, respectively. The experimental data set was shown in Table 2. The sample length under each dataset is 1024 sampling points, so the number of samples in both source domain and target domain is 1539. The training set and verification set are divided according to the ratio of 1:4, and the seed ratio of random number is 0.2. Therefore, the sample number of training set is 1231, and the sample number of verification set is 308.

Table 2. Classification of bearing faults

Damage location	Rolling element			Inner ring			Outer race			None
Tag of label	BF7	BF14	BF21	IF7	IF14	IF21	OF7	OF14	OF21	NC
Diameter with damage	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021	0

3.2 Network Training

With 256 size as one batch, the maximum number of the dataset epoch was set to 300 rounds, the optimizer is set to Adam and the initial learning rate (lr) was 0.001. Although the lr could be updated according to the fixed interval length, the lr of different intervals could not be updated diversified. Therefore, multi-step attenuation is used to realize dynamic interval length control.

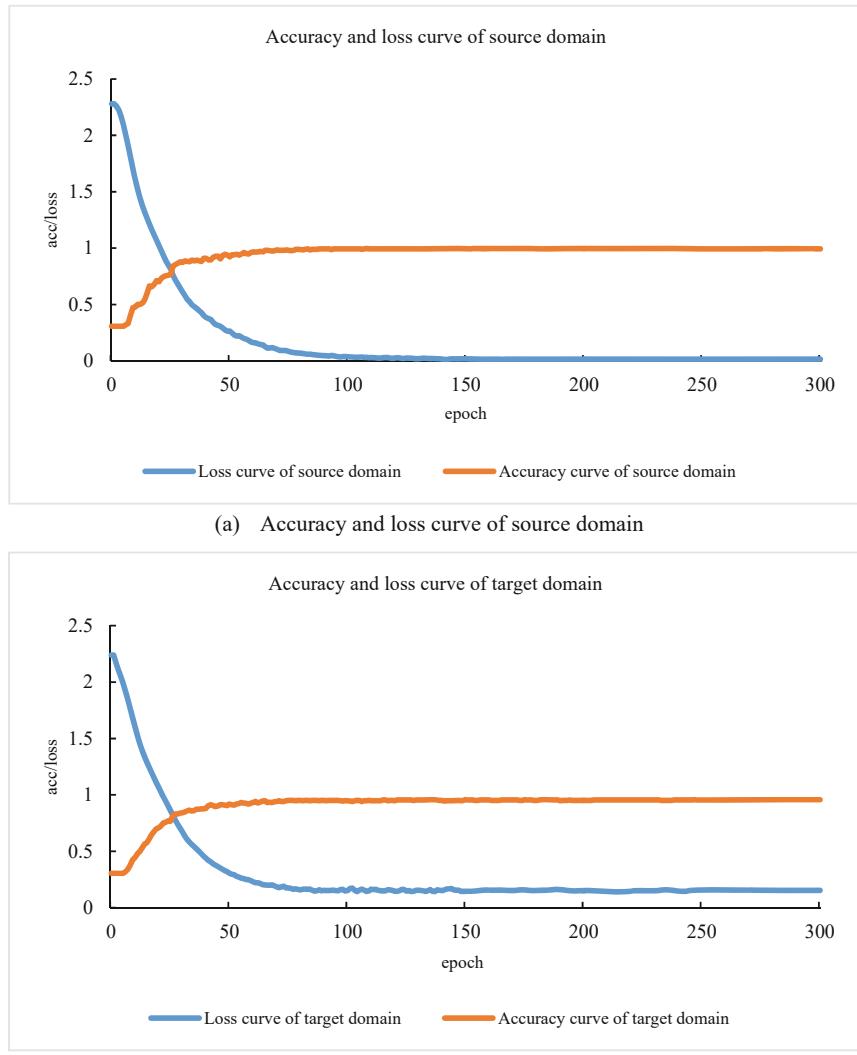
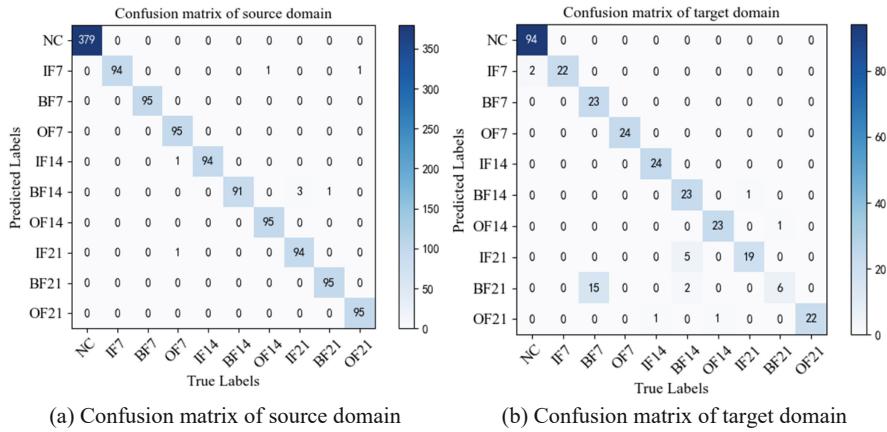


Fig. 4. Accuracy and loss curves

The preprocessed data is used as the input of 1D CNN. The experimental platform configuration is as follows: Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz, windows10 64-bit operating system, as well as the program running environment is Pytorch. As shown in Fig. 4, accuracy and loss curves of the training set and testing set after the training is completed. It can be found from Fig. 4, that the accuracy of source domain and target domain respectively reaches 100% and 99.5%, wherein the loss value of source domain and target domain tends to 0 after iteration 145 times. The value of accuracy has more than 95% accuracy after 100 iterations, and the model tends to be stable after 80 iterations.



(a) Confusion matrix of source domain

(b) Confusion matrix of target domain

Fig. 5. Confusion matrix of source domain and target domain

The confusion matrix is used to visualize the classification results after 300 iterations, as shown in Fig. 5. Figure 5 (a) is the confusion matrix of the source domain, the other is the confusion matrix of the target domain. The coordinate axis in the figure respectively represent the types of predicted and actual fault types. The results show that 1D CNN has separated the samples of different failure types well: The accuracy of source domain is nearly 100%, the accuracy of target domain is about 87.9%. The sensitivity of NC, BF7 and BF14 faults is a little worse. Double IF7 fault samples are wrongly classified into NC class, and five IF21 fault samples are wrongly classified into BF14 class. There are fifth BF21 fault samples that are wrongly assigned to the BF7 and five BF21 fault samples that are wrongly assigned to the BF14. However, there is a good classification effect for every other kind of fault.

3.3 Feature Visualization

In order to demonstrate the feature learning effect of 1D CNN based DTL in CWRU standard data set, t-distribution Stochastic Neighbor Embedding (t-SNE) [12, 13] technology is applied to visualize the features extracted from 1D CNN output layer, as shown in Fig. 6 and Fig. 7, where 10 different colors represent 10 different bearing fault types. In Fig. 6, 10 types of different faults in the data set are intermixed, making it difficult to

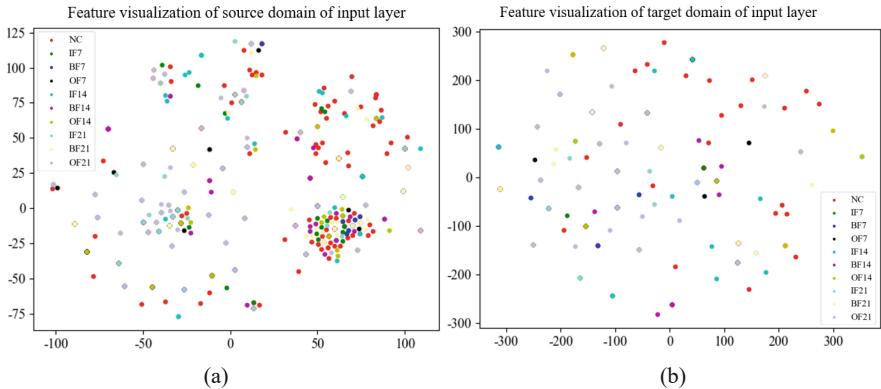


Fig. 6. T-SNE dimensionality reduction visualization of input layer

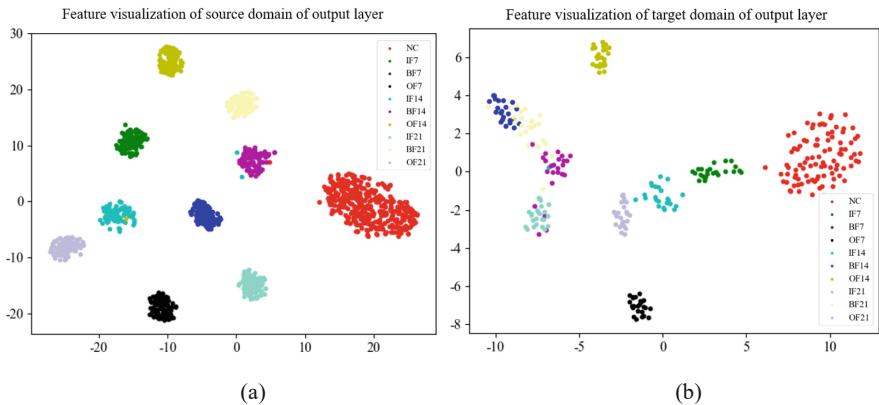


Fig. 7. T-SNE dimensionality reduction visualization in output layer

distinguish different bearing fault types. In Fig. 7, the source domain and target domain are pre-processed and put into the 1D CNN. In addition, t-SNE is used to cluster inputting layer and outputting layer, and the same fault types are clustered together. Therefore, t-SNE feature visualization results show that 1D CNN has high recognition accuracy and fault type classification ability.

3.4 Performance Comparison

Table 3. Summary of classifications' accuracy

Experimental method	Accuracy					Average
ResNet	73.62	73.04	73.35	73.61	73.15	71.669
	72.95	72.16	74.01	71.62	75.95	
	69.16	68.43	69.68	76.13	69.63	
	66.18	69.15	70.65	70.93	69.98	
Model migration	88.32	89.41	88.93	88.56	89.72	89.324
	88.51	89.46	88.06	90.01	90.05	
	88.43	88.16	89.63	89.58	89.36	
	89.46	89.87	89.85	90.65	90.46	
TCA	20.64	18.65	18.48	19.46	19.05	89.324
	18.26	18.55	18.19	20.47	20.84	
	19.23	20.14	19.15	20.01	19.48	
	20.27	19.62	18.59	18.47	19.93	

In order to verify the effectiveness of model migration fault diagnosis method, a comparative test was conducted with ResNet convolutional neural network and TCA diagnosis method. In order to reduce the impact of randomness on diagnosis results, three different experimental methods were divided into a group for each classification, and each group of tests was repeated 20 times. The classification accuracy is shown in Table 3. The comparison of classification accuracy is shown in Fig. 8.

By comparing the experimental results of different fault diagnosis methods, the average diagnosis accuracy of the proposed method based on model migration is higher than that of other methods. Firstly, due to the significant difference in the probability density distribution of data under different working conditions in the CWRU data set, the diagnosis ability of neural network is weakened, which leads to the low diagnosis accuracy of ResNet convolutional neural network diagnosis method. The TCA method has low diagnostic accuracy due to its lack of distribution adaptation to this process.

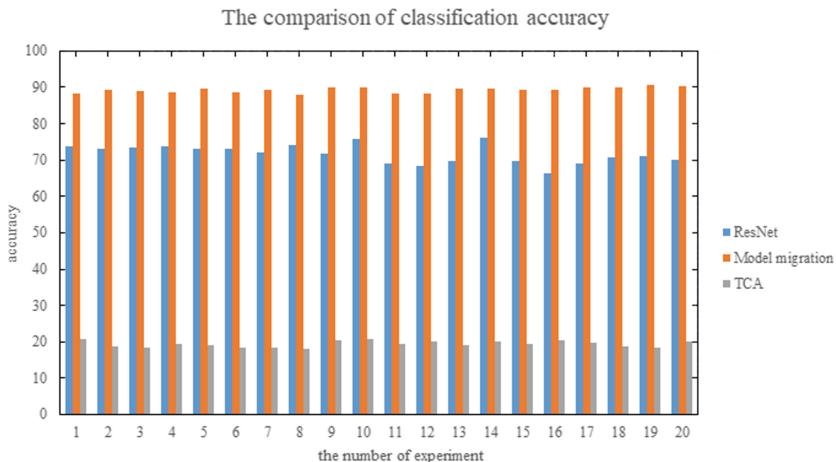


Fig. 8. Comparison of accuracy

4 Conclusion

In this paper, using model migration to extract the data sample of high suitability and CNN characteristics of high efficiency, it studied the fault diagnosis method based on the model of migration, and is verified by CWRU dataset, under different working conditions, to achieve the mechanical equipment of rolling bearings adaptive classification, realized the end to end of intelligent fault diagnosis, the main conclusions are as follows:

- (1) The model migration adaption to the same distribution samples, and CNN has the advantage of efficiently extracting sample features and updating model parameters. After 300 iterations, the feature classification accuracy of extracted target domain and source domain is close to 95%. Therefore, compared with traditional methods, the proposed method has significantly improved the extraction effect of fault features under different working conditions.
- (2) The transfer learning algorithm of different network models was used for comparative analysis, namely, ResNet convolutional neural network and TCA classification. The accuracy of bearing fault diagnosis of the proposed method reached 87.9%, higher than the other two methods. Therefore, compared with other methods, the proposed method reduces the dependence on labels and data samples and improves the accuracy of diagnosis.

Acknowledgement. This research is a part of the research that is sponsored by the Wuhu Science and Technology Program (No. 2021jc1-6).

References

1. Kang, S., Qiao, C., Wang, Y., et al.: Fault diagnosis method of rolling bearings under varying working conditions based on deep feature transfer. *J. Mech. Sci. Technol.* **34**(11), 4383–4391 (2020)
2. Lei, Y., Yang, B., Du, Z.: Deep transfer diagnosis method for machinery in big data era. *J. Mech. Eng.* **55**(7), 1–8 (2019)
3. Shao, H., Zhang, X., Cheng, J.: Intelligent fault diagnosis of bearing using enhanced deep transfer auto-encoder. *J. Mech. Eng.* **56**(9), 84–90 (2020)
4. Zhuang, F., Luo, P., He, Q.: Survey on transfer learning research. *J. Softw.* **26**(1), 26–39 (2015)
5. Shen, F., Chen, C., Yan, R.: Application of SVD and transfer learning strategy on motor fault diagnosis. *J. Vib. Eng.* **30**(1), 118–126 (2017)
6. Chen, C., Shen, F., Yan, R.: Enhanced least squares support vector machine-based transfer learning strategy for bearing fault diagnosis. *Chin. J. Sci. Instrum.* **38**(1), 33–40 (2017)
7. Paul, V., Meinecke, F., Klaus-Robert, M.: Finding stationary subspaces in multivariate time series. *Phys. Rev. Lett.* **103**(21), 214101 (2009)
8. Shi, Y., Sha, F.: Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In: Proceedings of the 29th International Conference on International Conference on Machine Learning, pp. 1275–1282 (2012)
9. Gong, B., Grauman, K., Fei, S.: Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2066–2073 (2012)
10. Pan, S., Tsang, I., Kwok, J.: Domain adaptation via transfer component analysis. *IEEE TNN* **22**(2), 199–210 (2011)
11. Smith, W.A., Randall, R.B.: Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **64–65**, 100–131 (2015)
12. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
13. Laurens, V., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11), 2579–2605 (2008)



Yak Management Platform Based on Neural Network and Path Tracking

Yunfan Hu^(✉)

University of Electronic Science and Technology of China, Chengdu, China
867822653@qq.com

Abstract. Positioning technology including path tracking is widely used in animal husbandry nowadays to help farmers to manage the animals and take care of them. At present, many methods have been utilized to track animals, distinguish or infer the behaviours of the animals, but there is still no complete and convenient platform that gives farmers clear and useful information about the animals. Besides, as for managing the animals who live in hard environment, for example, yaks, some difficulties may come from the challenging terrain. In this paper, GPS is used to obtain the time space information of the yaks, which can be used to analyse the behaviours of the yaks along with neural network in machine learning, and the design and implementation of a management platform that helps farmers manage the yaks based on subareas are also included.

Keywords: GPS · Neural network · Animal management

1 Introduction

Many fields in human's life benefit from technology nowadays and farming is no exception. To the perspective of farmers, knowing the locations and movements of animals helps a lot in managing or even taking good care of them. Positioning devices can record the location of animals and acknowledge their movement, which related to their behaviours like hunting, wandering, or doing other activities. With the information, the farmers can get familiar with the habits and living status of animals. Yaks are animals who live at high altitudes with steeps, sometimes it is challenging to set sensors and cameras there to monitor them and help farming. Therefore, method based on long-distance positioning technology should be figured out to handle such tasks.

Since late 1950s, researchers have paid attention to use telemetry tracking systems to learn about animals through collars or tags, however, the initial methods were costly and inaccurate [1]. In these years, as the spread of Global Positioning System (GPS), many researchers have utilized this satellite-based positioning system to track the location of animals to obtain the movements of them and solve specific problems. Chris J. Johnson et al. did a great number of experiments to find out the factors that influence the performance of GPS collars [2]. Patrick E. Clark et al. built an animal tracking system with a GPS tracking collar and a portable base station which can communicate wirelessly so that the base station can collect real-time location of the collar [1]. Vishwas Raj Jain

et al. built a Wireless Sensor Network (WSN) system based on GPS, which can collect different kinds of information, including location, temperature, head orientation and so on, and attempted to track swamp deer [3].

Only knowing the location is not enough to manage animals well, extracting useful information like animal habits or routines from the geographical data is also quite significant for animal management. Eugene D. Ungar et al. used GPS collars to continuously collect the location of cattle and classified their behaviours such as grazing and resting through regression method [4]. Petra Löttker et al. found dual-axis acceleration sensors in GPS collars help a lot in distinguishing animal behaviours because of knowing the accelerations in x and y axis [5].

During the past decade, researchers have introduced machine learning into inferring creaturel behaviours as well. Zhen Qin et al. proposed a deep neural network architecture that can recognize human activities based on multiple sensor data [6]. Ella Browning et al. trained deep learning models to predict the diving activities of seabirds with information from GPS devices and found these models performed better than other predictive models like hidden Markov models [7]. Guiming Wang even compared the main principals and the applications of unsupervised and supervised learning with some example algorithms, such as hidden Markov models, state space models, random forests in inferring animal behaviours [8].

From the previous studies, many approaches have been proposed to track animals and infer animal behaviours. However, these studies only give the method and part of the solution and there is no complete and convenient real-time managing platform for farmers. Besides, the living environment of yaks, plateau, is a kind of challenging terrain and may cause new difficulties in management. Therefore, with such application scenario, the challenges of proposing management platform in this paper involve several aspects. Firstly, the platform should be transplantable to make sure that it can be used by different kinds of devices, such as computer, mobile phones and embedded devices. Besides, since different devices have different computing powers and storage capabilities, lightweight is also a significant feature of the platform to make sure that it can satisfy the requirements of different devices.

Considering the problems and challenges above, this paper is expected to contain the content of both the design and the implementation of a platform or a software that can help farmers manage the yaks in their farms, based on positioning technology. It uses animal tracking collars that based on Global Positioning System (GPS) to get the location of each yak, and then do some operations to help the keepers manage the yaks. This approach only needs to attach devices such as tracking collars on the yaks, which avoids setting large equipment on challenging terrain.

2 Theoretical Basis

2.1 Global Positioning System (GPS)

The Global Positioning System (GPS) is an American satellite-based navigation system, and it uses radio waves to transmit information between its users and the satellites [9]. GPS receiver who is located on or near the earth and have a line of sight to at least four GPS satellites is able to calculate both geolocation and time information of itself with

the data from the GPS satellites, and because of some geometrical reasons, the solution of GPS is exist and unique [10]. In this paper, a GPS tool is used to obtain the latitudes, the longitudes and time information of the yaks.

2.2 Neural Network (NN)

Machine learning is a series of algorithms that make machines simulate or achieve learning like human and then make some improvements. Machine learning has a huge category but can be generally classified into four main classes: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [11]. With the development of computing power, Neural Network (NN) has become a useful tool in machine learning and gained its popularity because of its ability to distinguishing features with accuracy [12–14].

There are many kinds of neural networks, however, in this paper, the platform is used for operating the real-time information of the yaks and providing feedback to users as soon as possible, so the training and processing time is an important criterion. Compared to other neural network such as CNN with convolutional layer and pooling layer, basic NN has laconic architecture which is able to save time and increase efficiency.

In the following sections of this paper, neural network will be used to build a function that can classify the yaks according to their locations. Each class represents a subarea of the whole yak habitat, so the classified yaks are placed into different subareas. In this case, the management platform can do further operation to the information of yaks, including counting the numbers of yaks in each subarea, inferring the behaviours of yaks, detecting the issues when necessary and sending warnings to farmers. A neural network consists of several layers of processing nodes, which can do computations in parallel [15].

As for training the NN of this platform, the training data is the input of the first layer, which is called input layer. Then, few layers are used to operate the data and extract useful features of the data. The last layer is called output layer, its output should be the final result of the NN. This output should then be compared to the desired output to check the effectiveness of NN, and the difference between real output and the desired output, also known as the loss function, is the criterion for adjusting the elements of the transform matrices among the layers. After thousands of training epochs, the elements of the matrices will be adjusted to appropriate values and allow the NN obtain the output that very close to the desired one, which means the NN has been trained successfully. For using the NN after training, the real information should pass through the trained NN and the output will be obtained from output layer.

The structure of the layers and the nodes in neural network is linear, so activation functions are needed to perform nonlinear computations in specific situations. One of the most popular activation functions for classification is the sigmoid function [16], and this (1) will be used in all the layers except from the output layer.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

In terms of the output layer, the activation function should be chosen according to the output form. In this paper, the output of NN should represent the subarea labels, and the progress of passing the NN is to find the most likely subarea where the yak is located in, so using the output that represents the probability of each subarea is straightforward and suitable. The softmax function is able to normalize the output to probability and is widely used in machine learning nowadays [17]. The following Eq. (2) shows the softmax function.

$$P(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad \text{for } j = 1, 2, 3 \dots K \quad (2)$$

With the softmax function, the output of the NN is a vector whose length is the number of subareas and each element indicates the probability of corresponding subarea. Then a simple algorithm is performed to find the subarea with the largest probability and regard it as the exact one where the yak is located in. In this case, the NN can be firstly trained by the training data, and then used to find the subarea number of the yaks with real geographical information.

3 Research Content

3.1 Yak Counting Based on Subareas

For managing the yaks from a long distance, the positioning information of the yaks is quite significant, farmers need to know where the yaks are and how many yaks are doing certain activities. This makes farmers get familiar with the routines and habits of the yaks, so that they can make some decisions or some changes to management strategy. Therefore, the platform should be able to count the numbers of the yaks in all the subareas. The first step of this is to build and train a neural network that can distinguish all the subareas and place all the yaks into subareas according to their real-time location. The input of the NN should be the latitudes and the longitudes of the yaks and the output should be the probabilities of the yaks locating in corresponding subareas. The sizes of the transform matrices among all the layers should satisfy the scales of input and output, and the values of the elements in the transform matrices should be randomly initialized and then be adjusted during training. After passing through the successfully trained NN, the subareas where the yaks are located in are already known, and the yaks should be labelled with the subarea numbers that obtained in the previous progress to make it convenient to do further operations. With the labelled yaks' information, counting the number of yaks for each subarea becomes quite easy, it just needs to calculate the numbers of each subarea label. This can be achieved by performing an algorithm which scans all the labels of the yaks and keep a record of the total numbers of the subarea numbers.

3.2 Individual Analysis Based on Path Tracking

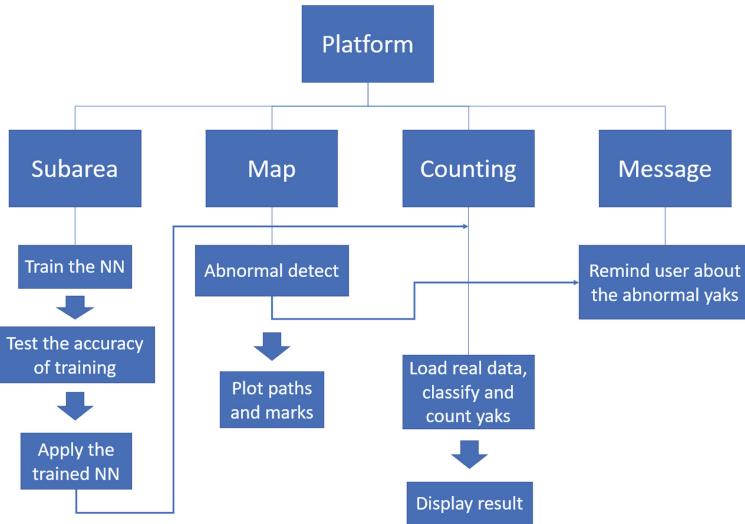
Only knowing the real-time positions of the yaks is not adequate to infer the exact status and behaviours of them, so the movement of the yaks should also be found and used to help farmers to infer the behaviours of the yaks. Fortunately, GPS also returns the time information along with the location, so the displacements between every pair of sample points of the same yak can be calculated and regarded as the movements of this yak. Considering these, the platform will calculate the displacements among the sample points of each yak, and check the movements of the yaks according them, detect whether some of the yaks have abnormal behaviours or not. Then the platform will send warnings to the farmer and plot the paths and real-time positions of the abnormal yaks to help farmer to check or take care of them. This requires the serial number of each yak, and their positions in a certain time interval. Therefore, compared to the data collected in previous section, the data collected for this function need the unique number of each yak to distinguish them, and the time along with the longitudes and the latitudes. After collecting the data, arrangement that number the yaks, list the positioning information according to timeline is necessary since the data are received randomly and may not in order. Besides, for the sample points of each yak, sorting the longitudes and the latitudes according to the time also makes it easy to draw the paths of the yak, which will be used in inferring behaviours and detecting abnormal yaks.

In order to know which yaks are abnormal, a detecting algorithm should be performed. For example, the yaks who stay at one place during an unreasonable time period and for long time is regarded as an abnormal one, but considering the error of positioning technology, a threshold value should be set to be the standard of “stay at one place”. For this circumstance, if all the displacements among the positions of the yak are smaller than the threshold, and the total time span is larger than another threshold, the yaks is judged to be abnormal and the sent warning should include the number of the yak. In fact, many other situations can be regarded as a mishap occurs, for instance, a yak stay outside the sleep area when other yaks are sleeping, or a yak stay at sleeping place when others are doing activities outside. All the situations similar to these should be detected, and this can be achieved by comparing the positions and paths of the yaks with their routines, or comparing individuals with the trends of yak herd. After inferring the behaviours of yaks and detecting abnormal situations, to help farmers monitor the movements of the yaks and know the positions of the abnormal yaks, the platform should plot the paths and latest positions of the yaks on the map and use different marks to distinguish the abnormal yaks and others.

4 Function Design and Implementation

4.1 Functions

See Fig. 1.

**Fig. 1.** System diagram.

a) Subareas and Virtual fences

The platform should be able to divide the whole living field into multiple subareas according to the latitude and the longitude. To achieve this function, neural network is proposed to deal with the positions of the yaks and get the subarea information of them as the output. The architecture of the neural network should be defined according to the sizes of inputs and the outputs.

In this paper, each of the longitudes and the latitudes collected by GPS tool contains totally 8 valid digits for integral part and fractional part. To make sure that enough features can be extracted and all the digits can be treated equally, the inputs of the neural network are defined to be 16-element vectors and the elements are the valid digits of longitudes and latitudes. Therefore, the input layer should have 16 neurons to process each element of the input. As for the output layer, since the softmax function allows output to represent the probabilities of the subareas, the number of neurons should be equal to the number of the subareas, which is 5 in this paper. Apart from input layer and output layer, two hidden layers are included between them to avoid failure in Linear inseparable circumstances, and the numbers of neurons of them are 12 and 9 in this paper.

After building the NN, the training data which contains a great number of sample points with their latitude, longitude and the serial number of the subarea, also called the subarea label, should be inputted to the randomly initialized NN. Then, during the training, the NN should adjust its transform matrices among the layers according to the differences between its real outputs and the desired outputs, which are the subarea labels of the sample points in the training data. After the training, the NN can be used to perform classification function with the real-time positions of the yaks and label the yaks with their subarea number.

As for the virtual fences, they can be clearly seen in the plot of the sample points. To be more specific, the borders of the subareas are the virtual fences, and they can be roughly the connecting lines of the sample points that are located near the borders of subareas if the number of sample points are large enough. If necessary, some algorithms that informs the farmers or makes the GPS collars shaking when yaks get close to the subarea borders can be added to make practical use of the virtual fences.

b) Counting

After labelling the yaks with their subarea numbers, counting the number of yaks for each subarea becomes quite simple. The platform should scan all the subarea labels, and keep a record of the total number of the labels with the same subarea number. Then, the result should be displayed on the graphical user interface (GUI) of the management platform.

c) Warning

Sometimes mishaps may happen to the animals, for example, sickness, death, being hurt by others during animal conflicts and so forth. In this case, the platform has to send warnings to the farmers to inform them so that they can take first aid and take good care of the yaks. To achieve this function, the behaviours of the yaks should be inferred from the time and positioning information from GPS. The platform builds individual files (shown in Fig. 2) for the yaks, which contains their positions and the subarea where they are located in along with time.

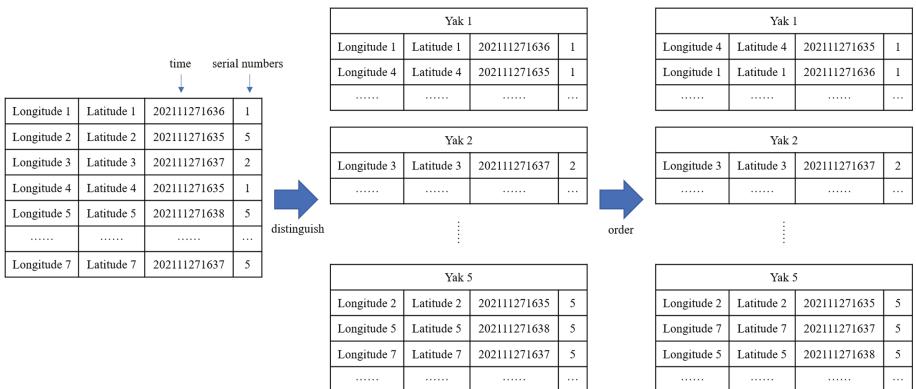


Fig. 2. Building individual files for yaks.

All the positions are sorted according to the timeline and the paths of yaks are plotted on the GUI to help farmers monitor the movements of the yaks and infer their behaviours. Then, with the paths and behaviours, an algorithm is performed to detecting the mishap. This algorithm detects the behaviours of individuals and compares them with most of the yak behaviours at the same time and the routines of the yaks, if there are some remarkable differences between them, the individual is judged to be the abnormal one. In this case, the algorithm can consider many kinds of specific situations and find

the abnormal yaks among the herd. To inform the farmers the serial numbers and the positions of the abnormal yaks and help farmers find them, the GUI will plot the positions of the abnormal ones with special marks, which can be easily distinguished from others.

4.2 Test of Functions

To check the effect of the platform, since all of the functions are based on the subarea classification, the accuracy of classification is a very important criterion. In terms of collecting data of the NN, both the training data and the testing data which contain the positioning information and subarea numbers of each yak should be collected. The subarea numbers in training data are used for obtaining the loss function and adjust the transform matrices of the NN, while the subarea number in testing data is used for calculating the accuracy of the classification.

As for the training data, 550 sample points (Fig. 3) are randomly chosen from a bounded region and classified into 4 special subareas, which means the parts with specific utilizations, and a common subarea (marked in blue), which means the other parts of the whole living environment. Then the training data is used one by one to train the NN, and after 10000 training epochs, the NN is successfully trained.

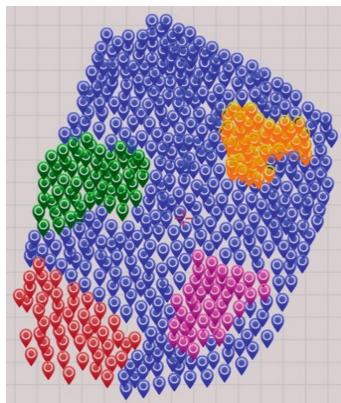


Fig. 3. The training data.

As for the testing data, several sets of 60-sample-point data are passed through the NN to check the accuracy of subarea classification and then the counting algorithm. In each set, there are 10 yaks in each special subarea and 20 yaks in the common subarea. Three sets of the sample points are shown below as examples, which is followed by the results of the classification (Figs. 4 and 5).

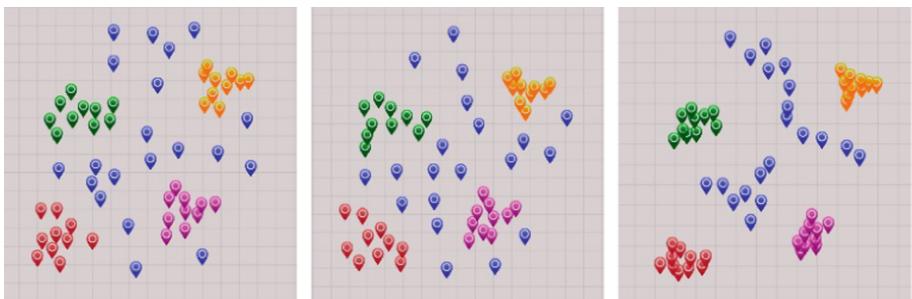


Fig. 4. Three sets sample points of the testing data.

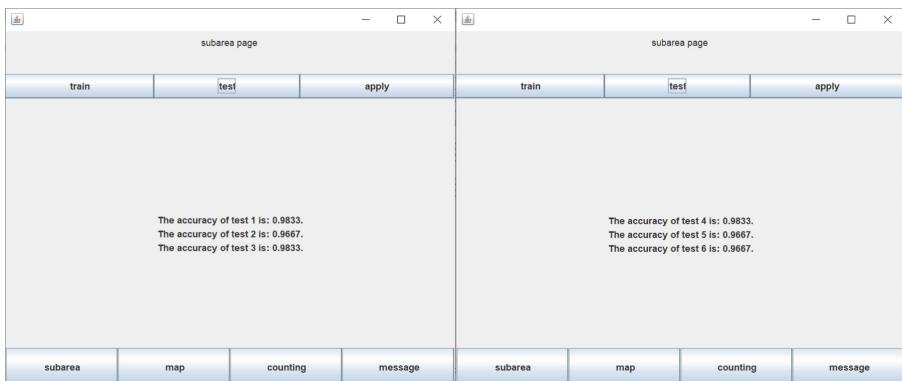
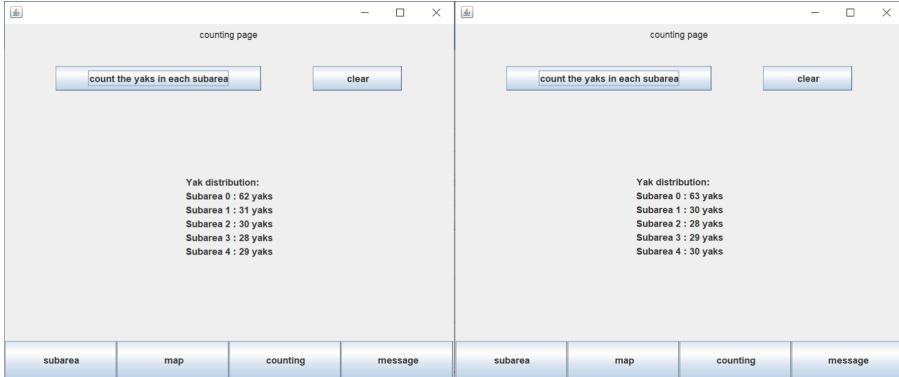


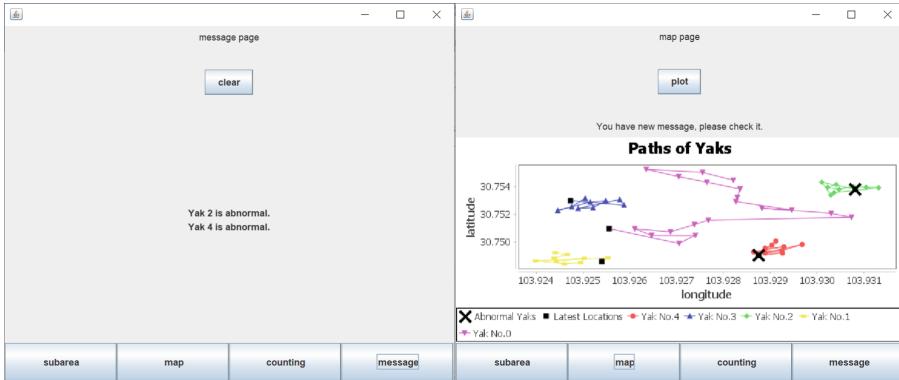
Fig. 5. The result of classification.

From the results above, the accuracies are all higher than 95% and acceptable in the scenarios of this paper. Same sets of data are also used for testing the effectiveness of CNN, and the results shows that compared to basic NN, CNN not only cannot increase the accuracy of classification, but also cost more time in training and classifying.

In terms of counting the yaks in each subarea, two sets of sample points are regarded as yaks to check the effectiveness of the counting algorithm. In each set, there are 60 points in the common area, whose serial number is 0, and there are 30 points in each special subarea. The results of counting are shown in the figure below (Fig. 6).

**Fig. 6.** The results of counting.

From the yak distribution obtained by the counting algorithm, few yaks are classified into wrong subarea. This is caused by the errors in classification, so the accuracy of counting is same as the one of classification, which has been already considered to be acceptable in previous paragraph. Then, to test the function of path drawing, abnormal detecting and warning sending, the positions of several yaks in a certain time period are collected along with the time. The testing results are shown below, and the abnormal yaks are marked differently (black cross) with others. According to Fig. 7, Yak 2 and Yak 4 are judged to be abnormal ones since the moving distances of both among 10 sampled time slots are too small for healthy yaks, so the abnormal detection can work well.

**Fig. 7.** Warnings and paths.

5 Conclusion

In this paper, a yak management platform based on positioning technology has been successfully designed and implemented. This platform has been able to classify the yaks

into different classes according to their position information that comes from GPS, count the number of yaks in each subarea of the living environment of yaks, draw the paths and send warnings to farmers when mishap happens. Compared to the existing related methods that only analysed the animals individually, the yak management platform in this paper has performed both individual analysis and a partition managing strategy that can infer the states and behaviours of herds, consider group activities when doing individual analysis, which has dealt with the problem from more perspectives. In addition, the platform has overcome the lack of complete long-distance management platform for animal keepers, but also has built a model structure to animal management software, which can be improved and extended into other scenarios and make more contributions in animal husbandry.

References

1. Clark, P.E., et al.: An advanced, low-cost, GPS-based animal tracking system. *Rangel. Ecol. Manage.* **59**, 334–340 (2006)
2. Johnson, C.J., Heard, D.C., Parker, K.L.: Expectations and realities of GPS animal location collars: results of three years in the field. *Wildl. Biol.* **8**, 153–159 (2002)
3. Jain, V.R., Bagree, R., Kumar, A., Ranjan, P.: wildCENSE: GPS based animal tracking system. In: International Conference on Intelligent Sensors, Sensor Networks and Information Processing (2008)
4. Ungar, E.D., Henkin, Z., Gutman, M., Dolev, A., Genizi, A., Ganskopp, D.: Inference of animal activity from GPS collar data on free-ranging cattle. *Rangel. Ecol. Manage.* **58**, 256–266 (2005)
5. Löttker, P., et al.: New possibilities of observing animal behaviour from a distance using activity sensors in GPS-collars: an attempt to calibrate remotely collected activity data with direct behavioural observations in red deer *Cervus elaphus*. *Wildl. Biol.* **15**, 425–434 (2009)
6. Qin, Z., Zhang, Y., Meng, S., Qin, Z., Choo, K.-K.R.: Imaging and fusing time series for wearable sensor-based human activity recognition. *Inf. Fusion* **53**, 80–87 (2020)
7. Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T., Freeman, R.: Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. *Methods Ecol. Evol.* **9**, 681–692 (2017)
8. Wang, G.: Machine learning for inferring animal behavior from location and movement data. *Ecol. Inform.* **49**, 69–76 (2019)
9. The Official of The Global Positioning System (GPS): What is GPS? www.gps.gov/systems/gps/. Accessed 16 Jan 2022
10. Abel, J.S., Chaffee, J.W.: Existence and uniqueness of GPS solutions. *IEEE Trans. Aerosp. Electron. Syst.* **27**, 952–956 (1991)
11. Richard Yu, F., He, Y.: Deep Reinforcement Learning for Wireless Networks. Springer, Cham (2019). <https://doi.org/10.1007/978-3-030-10546-4>
12. Ding, Y., et al.: MVFusFra: a multi-view dynamic fusion framework for multimodal brain tumor segmentation. *IEEE J. Biomed. Health Inform.* **26**, 1570–1581 (2021)
13. Qin, Z., Huang, G., Xiong, H., Qin, Z., Choo, K.-K.R.: A fuzzy authentication system based on neural network learning and extreme value statistics. *IEEE Trans. Fuzzy Syst.* **29**, 549–559 (2021)
14. Qin, Z., et al.: Learning-aided user identification using smartphone sensors for smart homes. *IEEE Internet Things J.* **6**, 7760–7772 (2019)

15. Wang, S., Zhang, X., Zhang, Y., Wang, L., Yang, J., Wang, W.: A survey on mobile edge networks: convergence of computing, caching and communications. *IEEE Access* **5**, 6757–6779 (2017). <https://doi.org/10.1109/ACCESS.2017.2685434>
16. Li, H., et al.: A novel feedrate scheduling method based on Sigmoid function with chord error and kinematic constraints. *Int. J. Adv. Manuf. Technol.*, 1–22 (2021). <https://doi.org/10.1007/s00170-021-08092-1>
17. Jap, D., Won, Y.-S., Bhasin. S.: Fault injection attacks on SoftMax function in deep neural networks. In: *Computing Frontiers* (2021)



Stability Analysis of Hopfield Neural Networks with Conformable Fractional Derivative: M-matrix Method

Chang-bo Yang^{1,2(✉)}, Sun-yan Hong³, Ya-qin Li¹, Hui-mei Wang¹, and Yan Zhu¹

¹ Institute of Nonlinear Analysis, Kunming University, Kunming 650214, Yunnan, People's Republic of China
cbyang348@126.com

² School of Physics and Astronomy, Yunnan University, Kunming 650500, Yunnan, People's Republic of China

³ School of Information Engineering, Kunming University, Kunming 650214, Yunnan, People's Republic of China

Abstract. In this work, the stability analysis of a class of conformable fractional-order Hopfield neural networks is investigated. By using the Lyapunov function and M-matrix method, certain novel results on the existence, uniqueness and fractional exponential stability of the equilibrium point have been established. The derived criteria improve and extend some recent results in the literature. Finally, the advantage of our theoretical results is illustrated via a numerical example.

Keywords: Hopfield neural networks · Conformable fractional derivative · M-matrix · Stability

1 Introduction

Hopfield Neural Networks (HNNs) [1], one of the famous recurrent neural networks (RNNs), has drawn considerable attentions of many scholars due to its extensive applications such as associative memory, pattern recognition and combinatorial optimization [2–4]. As we know, this original model is described by a group of differential equations with integeral first-order derivative. Compared to the integral derivative, fractional derivative can provide a better tool for the description of memory of neurons. In 2009, Boroomand and Menhaj [5] proposed a fractional-order HNNs by using a generalized capacitor and its stability is discussed by energy-like function method. This new model is successfully applied to optimization parameter estimation problem. After that, various fractional-order versions of HNNs have been established and their stability analysis have been carried out [6–11]. However, most of results are in the Riemann-Liouville sence

This research is partially supported by the Research Education-funded Projects in Yunnan Province (2021J0713), the Natural Science Foundation of Yunnan Provincial Department of Science and Technology (202101BA070001-132) and the Research Fund of Kunming University (YJL17004, YJL20015, YJL20019).

or the Caputo sence. These two types of fractional derivatives face some shortcomings such as the complexity of definitions and the difficulty in calculation. Recently, authors in [12] proposed a new class of HNNs with conformable fractional derivative, which was described by the following conformable differential equations:

$$\begin{cases} T_\alpha x_i(t) = -a_i x_i(t) + \sum_{j=1}^n b_{ij} f_j(x_j(t)) + I_i, & t > 0, \\ x_i(0) = x_{i0}, & i \in N = \{1, 2, \dots, n\}, \end{cases} \quad (1)$$

or compact matrix form:

$$\begin{cases} T_\alpha x(t) = -Ax(t) + Bf(x(t)) + I, & t > 0, \\ x(0) = x_0, \end{cases}$$

where $x_0 = (x_{10}, x_{20}, \dots, x_{n0})^T \in \mathbb{R}^n$ is the initial condition, $A = \text{diag}(a_1, a_2, \dots, a_n)$, $a_i > 0$ denote the passive decay rates, $B = (b_{ij})_{n \times n} \in \mathbb{R}^{n \times n}$ is the interconnection weight matrix of the network, $I = (I_1, I_2, \dots, I_n)^T \in \mathbb{R}^n$ is the external input vector, $f(\bullet)$ is the activation function, $T_\alpha x(t)$, $\alpha \in (0, 1]$ denotes the conformable derivative (see Definition 1) of the state vector.

With regard to model (1), some stability criteria are in the form of algebraic inequalities [12] and linear matrix inequality (LMI) [13] are established, respectively. To best our knowledge, the M-matrix is also an important tool in stability analysis of NNs [14–17]. In this paper, we shall further study the stability issue of model (1) by M-matrix theory and some analysis techniques.

To ensure the existence of the solution of model (1), normally, we assume that the activation function $f(\bullet)$ satisfies the following lipschitz condition:

(A1) $f(\bullet) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuous and there exists a positive diagonal matrix

$$L = \text{diag}(L_1, L_2, \dots, L_n)$$

such that $|f(x) - f(y)| \leq L|x - y|$ for any $x, y \in \mathbb{R}^n$.

2 Preliminaries

Definition 1. [18] The conformable fractional derivative of a given function $f(t)$ of order $0 < \alpha \leq 1$ is defined as follows:

$$T_\alpha f(t) = \lim_{\varepsilon \rightarrow 0} \frac{f(t + \varepsilon t^{1-\alpha}) - f(t)}{\varepsilon}, \quad \forall t > 0. \quad (2)$$

From Definition 1, for any $k_1, k_2 \in \mathbb{R}$, f, g are α -differentiable at $t > 0$, we are easy to get the following properties [19]:

(a) Linearity property

$$T_\alpha(k_1 f + k_2 g) = k_1 T_\alpha f + k_2 T_\alpha g, \quad (3)$$

(b) Product formula

$$T_\alpha f g = g T_\alpha f + f T_\alpha g, \quad (4)$$

(c) Quotient formula

$$T_\alpha \left(\frac{f}{g} \right) = \frac{g T_\alpha f - f T_\alpha g}{g^2}. \quad (5)$$

From the above properties, we can see that the conformable fractional derivative inherits a lot of good properties from integral-order derivative. In fact, if $f(t)$ is differentiable, then $T_\alpha f(t) = t^{1-\alpha} f'(t)$.

Definition 2. A constant vector $x^* = (x_1^*, x_2^*, \dots, x_n^*)^T \in \Re^n$ is called an equilibrium point of model (1) if

$$-Ax^* + Bf(x^*) + I = 0.$$

Definition 3. The equilibrium point $x^* \in \Re^n$ of model (1) is said to be fractional exponentially stable if

$$\|x(t) - x^*\|_2 \leq M \|x_0 - x^*\|_2 e^{-\lambda \frac{t^\alpha}{\alpha}}, \quad t > 0,$$

where the constants $M > 0$, $\lambda > 0$, and $x(t)$ is the arbitrary solution of model (1) with the initial condition x_0 .

Lemma 1. Considering the following conformable Cauchy problems:

$$\begin{cases} T_\alpha x(t) = -\lambda x(t) \\ x(0) = x_0 \end{cases}.$$

Then we have:

$$x(t) = e^{-\lambda \frac{t^\alpha}{\alpha}}, \quad \lambda > 0, \quad t > 0.$$

Proof: Let $y(t) = x(t)e^{\lambda \frac{t^\alpha}{\alpha}}$, using the property (b), we have:

$$\begin{aligned} T_\alpha y(t) &= T_\alpha x(t)e^{\lambda \frac{t^\alpha}{\alpha}} + x(t)\lambda e^{\lambda \frac{t^\alpha}{\alpha}} \\ &= -\lambda x(t)e^{\lambda \frac{t^\alpha}{\alpha}} + \lambda x(t)e^{\lambda \frac{t^\alpha}{\alpha}} \\ &= 0. \end{aligned} \quad (6)$$

Integrating both side of (6), we have $y(t) - y(0) = 0$. This is

$$y(t) = y(0) \Rightarrow x(t) = x_0 e^{-\lambda \frac{t^\alpha}{\alpha}}, \quad \lambda > 0, \quad t > 0.$$

Lemma 2. [20] Let $S \in Z^{n \times n}$, where $Z^{n \times n}$ is a set of $n \times n$ matrices with non-positive off-diagonal elements, then S is an M-matrix if and only if there exists a positive vector $\xi = (\xi_1, \xi_2, \dots, \xi_n)^T \in \Re^n$ such that $S\xi > 0$ or $\xi^T S > 0$.

3 Main Results

In this section, we firstly introduce a generalized norm [16, 21], which is defined by

$$\|x\|_{\{\xi, \infty\}} = \max_{1 \leq i \leq n} \{\xi_i^{-1} |x_i|\}, \quad \xi = (\xi_1, \xi_2, \dots, \xi_n)^T > 0.$$

Clearly, $\|\bullet\|_{\{\xi, \infty\}}$ is a nature generalization of $\|\bullet\|_\infty$. In fact, let $\xi = (1, 1, \dots, 1)^T$, $\|\bullet\|_{\{\xi, \infty\}}$ turns into the usual maximum norm. Based on this generalized norm and the contracting mapping principle, an improved criterion on the existence and uniqueness of equilibrium point of model (1) has been established. And then, the fractional exponential stability of such equilibrium point is investigated by constructing Lyapunov function and applying the interesting properties of conformable fractional derivative.

Theorem 1. Under condition (A1), then the model (1) has a unique equilibrium point $x^* = (x_1^*, x_2^*, \dots, x_n^*)^T \in \mathfrak{N}^n$ if the following condition is fulfilled:

(A2): $A - |B|L$ is an M-matrix, where $A = \text{diag}\{a_1, a_1, \dots, a_n\}$, $B = (b_{ij})_{n \times n}$, $L = \text{diag}\{L_1, L_1, \dots, L_n\}$.

Proof: Suppose that $x^* = (x_1^*, x_2^*, \dots, x_n^*)^T \in \mathfrak{N}^n$, by Definition 2, we have

$$Ax^* = Bf(x^*) + I \Rightarrow x^* = A^{-1}Bf(x^*) + A^{-1}I.$$

Let $H(x) = A^{-1}Bf(x) + A^{-1}I$, to prove Theorem 1, it is equivalent to show that $H(x) = x$ has a unique fixed point. To end this, we shall prove that $H(x)$ is a contraction mapping from \mathfrak{N}^n to \mathfrak{N}^n . Since $A - |B|L$ is an M-matrix, from Lemma 2, there exists a positive vector $\xi = (\xi_1, \xi_2, \dots, \xi_n)^T \in \mathfrak{N}^n$ such that $(A - |B|L)\xi > 0$. This is

$$a_i\xi_i - \sum_{j=1}^n |b_{ij}|L_j\xi_j > 0 \Rightarrow \sum_{j=1}^n |b_{ij}|L_j\xi_j < a_i\xi_i \Rightarrow \xi_i^{-1}a_i^{-1} \sum_{j=1}^n |b_{ij}|L_j\xi_j < 1.$$

Denote

$$\theta = \max_{1 \leq i \leq n} \left\{ \xi_i^{-1}a_i^{-1} \sum_{j=1}^n |b_{ij}|L_j\xi_j \right\},$$

we have $0 < \theta < 1$. On the other hand, by the Definition of $\|\bullet\|_{\{\xi, \infty\}}$ and the condition (A1), for any $x, y \in \mathfrak{N}^n$, we have

$$\begin{aligned} \|H(x) - H(y)\|_{\{\xi, \infty\}} &= \max_{1 \leq i \leq n} \left\{ \xi_i^{-1} |h_i(x) - h_i(y)| \right\} \\ &\leq \max_{1 \leq i \leq n} \left\{ \xi_i^{-1} a_i^{-1} \sum_{j=1}^n |b_{ij}| L_j \xi_j \xi_j^{-1} |x_j - y_j| \right\} \\ &\leq \max_{1 \leq i \leq n} \left\{ \xi_i^{-1} a_i^{-1} \sum_{j=1}^n |b_{ij}| L_j \xi_j \right\} \max_{1 \leq j \leq n} \left\{ \xi_j^{-1} |x_j - y_j| \right\} \end{aligned}$$

$$= \theta \|x - y\|_{\{\xi, \infty\}}.$$

It follows from $0 < \theta < 1$ that $H(x)$ is a contraction mapping from \Re^n to \Re^n . Thus there exists a unique $x^* \in \Re^n$ such that $H(x^*) = x^*$, which implies that model (1) has a unique equilibrium point $x^* \in \Re^n$. The proof is completed.

Theorem 2. Under conditions (A1) and (A2), then the unique equilibrium point $x^* \in \Re^n$ of model (1) is fractional exponentially stable if the following condition is satisfied:

$$(A3): C - |B|L \text{ is an M-matrix, where } C = \text{diag}\{c_1, c_1, \dots, c_n\}, \quad c_i = 2a_i - \sum_{j=1}^n |b_{ij}|L_j, B = (b_{ij})_{n \times n}, L = \text{diag}\{L_1, L_1, \dots, L_n\}.$$

Proof: It follows from Theorem 1 that model (1) has a unique equilibrium point $x^* = (x_1^*, x_2^*, \dots, x_n^*)^T \in \Re^n$. Substituting $y_i(t) = x_i(t) - x_i^*$ into model (1) leads to

$$T_\alpha y_i(t) = -a_i y_i(t) + \sum_{j=1}^n b_{ij}[f_j(y_j(t) + x_j^*) - f_j(x_j^*)], \quad t > 0.$$

Since $C - |B|L$ is an M-matrix, from Lemma 2, there exists a positive vector $\xi = (\xi_1, \xi_2, \dots, \xi_n)^T \in \Re^n$ such that $\xi^T(C - |B|L) > 0$. This is

$$\left(2a_i - \sum_{j=1}^n |b_{ij}|L_j\right)\xi_i - \sum_{j=1}^n |b_{ji}|L_i\xi_j > 0.$$

Furthermore, we have

$$\left(2a_i - \sum_{j=1}^n |b_{ij}|L_j\right) - \sum_{j=1}^n |b_{ji}|L_i\xi_j \xi_i^{-1} > 0.$$

Consider the following Lyapunov function:

$$V(t) = \sum_{i=1}^n \xi_i y_i^2(t).$$

By Definition 1 and the property (b), taking the conformable derivative of $V(t)$ along the solutions of model (1), we have

$$\begin{aligned} T_\alpha V(t) &= \sum_{i=1}^n \xi_i 2y_i(t) T^\alpha y_i(t) \\ &= \sum_{i=1}^n \xi_i 2y_i(t) \left\{ -a_i y_i(t) + \sum_{j=1}^n b_{ij} [f_j(y_j(t) + x_j^*) - f_j(x_j^*)] \right\} \end{aligned}$$

It follows from (A1) that

$$T_\alpha V(t) \leq -2 \sum_{i=1}^n \xi_i a_i y_i^2(t) + \sum_{i=1}^n \sum_{j=1}^n |b_{ij}| L_j \xi_i (2|y_i(t)| |y_j(t)|)$$

Using the inequality $2|a||b| \leq a^2 + b^2$, we have

$$\begin{aligned} T^\alpha V(t) &= -2 \sum_{i=1}^n \xi_i a_i y_i^2(t) + \sum_{i=1}^n \xi_i \sum_{j=1}^n |b_{ij}| L_j (y_i^2(t) + y_j^2(t)) \\ &= \sum_{i=1}^n \left(-2a + \sum_{j=1}^n |b_{ij}| L_j \right) \xi_i y_i^2(t) + \sum_{i=1}^n \sum_{j=1}^n \xi_i |b_{ij}| L_j y_j^2(t) \\ &= \sum_{i=1}^n \left(-2a + \sum_{j=1}^n |b_{ij}| L_j \right) \xi_i y_i^2(t) + \sum_{i=1}^n \sum_{j=1}^n |b_{ji}| L_i \xi_j \xi_i^{-1} \xi_i y_i^2(t) \\ &= \sum_{i=1}^n \left(-2a + \sum_{j=1}^n |b_{ij}| L_j - \sum_{j=1}^n |b_{ji}| L_i \xi_j \xi_i^{-1} \right) \xi_i y_i^2(t) \\ &\leq -\mu \sum_{i=1}^n \xi_i y_i^2(t) \\ &= -\mu V(t), \end{aligned}$$

where

$$\mu = \min_{1 \leq i \leq n} \left\{ 2a_i - \sum_{j=1}^n |b_{ij}| L_j - \sum_{j=1}^n |b_{ji}| L_i \xi_j \xi_i^{-1} \right\} > 0.$$

By Lemma 1, we have

$$V(t) \leq V(0) e^{-\mu \frac{t^\alpha}{\alpha}}, \quad t > 0.$$

Furthermore, we have

$$\|x(t) - x^*\|_2 \leq M \|x_0 - x^*\|_2 e^{-\lambda \frac{t^\alpha}{\alpha}}, \quad t > 0,$$

where $\lambda = \frac{\mu}{2} > 0$, $M = \sqrt{\frac{\max_{1 \leq i \leq n} \{\xi_i\}}{\min_{1 \leq i \leq n} \{\xi_i\}}} > 0$. By Definition 3, we can derive that the equilibrium point x^* of model (1) is fractional exponentially stable. This completes the proof.

4 A Numerical Example

In this section, we shall give a concrete example to illustrate that results herein are less conservative than Theorem 1 and 2 in [12], respectively. Consider the following 2-dimensional conformable Hopfield neural networks:

$$\begin{cases} T_\alpha x(t) = -Ax(t) + Bf(x(t)) + I, & t > 0 \\ x(0) = x_0 \end{cases}, \quad (7)$$

where the activation function $f(x) = |x|$ and the parameters are

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0.5 & 0.5 \\ -0.6 & 0.3 \end{pmatrix}, \quad I = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (8)$$

It is easy to check that $L = \text{diag}\{1, 1\}$,

$$A - |B|L = \begin{pmatrix} 0.5 & -0.5 \\ -0.6 & 0.7 \end{pmatrix}, \quad C - |B|L = \begin{pmatrix} 0.5 & -0.5 \\ -0.6 & 0.8 \end{pmatrix}.$$

By taking $\xi = (1, 0.9)^T$, it follows from Lemma 2 that $A - |B|L$ and $C - |B|L$ are M-matrices. Then, the assumptions (A1–A3) are satisfied. From Theorem 2, we can conclude that the unique equilibrium point $(0, 0)^T$ of the model (7) with parameters (8) is fractional exponentially stable. However, according to Theorem 1 and 2 in [12] respectively, one has

$$a_1 = 1, \quad L_1 b_{11} + L_1 b_{21} = 1.1, \quad b_{11} L_1 + b_{12} L_2 = 1.$$

Noting that

$$a_1 \not> L_1 b_{11} + L_1 b_{21}, \quad a_1 \not> b_{11} L_1 + b_{12} L_2.$$

So criteria obtained in [9] are invalid for the stability of model (7) with parameters (8). Meanwhile, to verify the correctness of our results, some computer simulations are also given in Fig. 1, where the fractional-order: $\alpha = 0.6$.

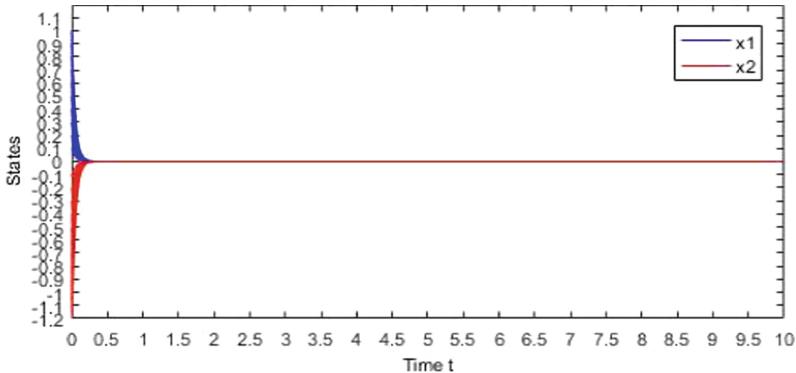


Fig. 1. State trajectories of $x(t) = (x_1, x_2)^T$ of the model (7) with parameters (8) and the initial conditions are $x_0 = (-0.1m, 0.1m)^T$, $m = 1, 2, \dots, 10$.

5 Conclusions

In this brief, stability analysis such as the existence, uniqueness and fractional exponential stability on equilibrium point to a class of conformable Hopfield neural networks has been further studied. Some new criteria in terms of M-matrix are obtained. Finally, an example is given to show the less conservative for our results. Objectively, time delays exist in the hardware implementation of networks. In the future, we shall discuss the stability of the model (1) with time delays.

References

1. Hopfield, J.J.: Neural network and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. **79**, 2554–2558 (1982)
2. Hopfield, J.J.: Neural computation of decisions in optimization problems. Biol. Cybern. **52**, 141–152 (1985)
3. Bouzerdoum, A., Pattison, T.: Neural networks for quadratic optimization with bound constraints. IEEE Trans. Neural Netw. **4**, 293–303 (1993)
4. Gao, W.X., Mu, X.Y., Tang, N., Yan, H.L.: Application of Hopfield neural network in unit commitment problem. J. Comput. Appl. **29**(4), 1028–1031 (2009)
5. Boroomand, A., Menhaj, M.B.: Fractional-order Hopfield neural networks. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008. LNCS, vol. 5506, pp. 883–890. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02490-0_108
6. Xia, H., Zhen, W., Li, Y.: Nonlinear dynamics and chaos in fractional-order Hopfield neural networks with delay. Adv. Math. Phys. **2013**, 1–9 (2013)
7. Wang, H., Yu, Y., Wen, G., Zhang, S.: Stability analysis of fractional-order neural networks with time delay. Neural Proc. Lett. **42**, 479–500 (2015)
8. Zhang, S., Yu, Y., Wang, H.: Mittag-Leffler stability of fractional-order Hopfield neural networks. Nonlinear Anal. Hybrid Syst **16**, 104–121 (2015)
9. Wang, H., Yu, Y., Wen, G., Zhang, S., Yu, J.: Global stability analysis of fractional-order Hopfield neural networks with time delay. Neurocomputing **154**, 15–23 (2015)

10. Chen, L., Liu, C., Wu, R., He, Y., Chai, Y.: Finite-time stability criteria for a class of fractional-order neural networks with delay. *Neural Comput. Appl.* **27**(3), 549–556 (2015). <https://doi.org/10.1007/s00521-015-1876-1>
11. Wu, H., Zhang, X., Xue, S., Niu, P.: Quasi-uniform stability of Caputo-type fractional-order neural networks with mixed delay. *Int. J. Mach. Learn. Cybern.* **8**(5), 1501–1511 (2016). <https://doi.org/10.1007/s13042-016-0523-1>
12. Kütahyaloglu, A., Karakoc, F.: Exponential stability of Hopfield neural networks with conformable fractional derivative. *Neurocomputing* **456**(6), 263–267 (2021)
13. Huyen, N.T.T., Sau, N.H., Thuan, M.V.: LMI conditions for fractional exponential stability and passivity analysis of uncertain Hopfield conformable fractional-order neural networks. *Neural Proc. Lett.* **54**(2), 1333–1350 (2022)
14. Yang, C., Zhou, X.-W., Wang, T.: Further analysis on stability for a class of neural networks with variable delays and impulses. In: Huang, D.-S., Gupta, P., Wang, L., Gromiha, M. (eds.) ICIC 2013. CCIS, vol. 375, pp. 13–18. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39678-6_3
15. Yang, C., Huang, T.: New results on stability for a class of neural networks with distributed delays and impulses. *Neurocomputing* **111**, 115–121 (2013)
16. Yang, C., Huang, T.: Improved stability criteria for a class of neural networks with variable delays and impulsive perturbations. *Appl. Math. Comput.* **243**, 923–935 (2014)
17. Liu, P., Zeng, Z., Wang, J.: Multiple Mittag-Leffler stability of fractional-order recurrent neural networks. *IEEE Trans. Syst. Man Cybern. Syst.* **47**(8), 2279–2288 (2017)
18. Khalil, R., Horani, M., Yousef, A., Sababheh, M.: A new definition of fractional derivative. *J. Comput. Appl. Math.* **264**, 65–70 (2014)
19. Atangana, A., Baleanu, D., Alsaedi, A.: New properties of conformable derivative. *Open Mathematics* **13**(1), 1–10 (2015)
20. Horn, R., Johnson, C.: Topics in Matrix Analysis. Cambridge University Press, London (1991)
21. Chen, T., Wang, L.: Global μ -stability of delayed neural networks with unbounded time-varying delays. *IEEE Trans. Neur. Netw.* **18**, 1836–1840 (2007)



Artificial Neural Networks for COVID-19 Forecasting in Mexico: An Empirical Study

C. M. Castorena¹ , R. Alejo² , E. Rendón² , E. E. Granda-Gutiérrez³ , R. M. Valdovinos⁴ , and G. Miranda-Piña²

¹ Tecnologías de la Información y las Comunicaciones, Universidad de Valencia, Av. Universitat s/n, 46100 Burjassot, Valencia, Spain

² Division of Postgraduate Studies and Research, National Institute of Technology of Mexico, (TecNM) Campus Toluca, Av. Tecnológico s/n, Agrícola Bellavista, 52149 Metepec, Mexico
MM22280266@toluca.tecnm.mx

³ UAEM University Center at Atlacomulco, Universidad Autónoma del Estado de México, Carretera Toluca-Atlacomulco Km. 60, 50450 Atlacomulco, Mexico

⁴ Faculty of Engineering, Universidad Autónoma del Estado de México, Cerro de Coatepec S/N, Ciudad Universitaria, 50100 Toluca, Mexico

Abstract. Artificial Neural Networks (ANN) have encountered interesting applications in forecasting several phenomena, and they have recently been applied in understanding the evolution of the novel coronavirus COVID-19 epidemic. Alone or together with other mathematical, dynamical, and statistical methods, ANN help to predict or model the transmission behavior at a global or regional level, thus providing valuable information for decision-makers. In this research, four typical ANN have been used to analyze the historical evolution of COVID-19 infections in Mexico: Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) neural networks, and the hybrid approach LSTM-CNN. From the open-source data of the Resource Center at the John Hopkins University of Medicine, a comparison of the overall qualitative fitting behavior and the analysis of quantitative metrics were performed. Our investigation shows that LSTM-CNN achieves the best qualitative performance; however, the CNN model reports the best quantitative metrics achieving better results in terms of the Mean Squared Error and Mean Absolute Error. The latter indicates that the long-term learning of the hybrid LSTM-CNN method is not necessarily a critical aspect to forecast COVID-19 cases as the relevant information obtained from the features of data by the classical MLP or CNN.

Keywords: COVID-19 · Forecasting · Artificial Neural Networks · Deep learning

This work has been partially supported under grants of project 11437.21-P from TecNM and 6364/2021SF from UAEMex.

1 Introduction

Artificial Neural Networks (ANN) have become a hot topic in artificial intelligence, particularly Deep Learning ANN (DL-ANN), which have been successfully employed in the classification of images, audio, and text, among others [9]. In addition, the DL-ANN have shown remarkable effectiveness in approximation functions, and prediction or forecast [10]. In the recent pandemic occasioned by coronavirus disease (COVID-19), DL-ANN have confirmed the ability to forecast COVID-19 cases. Ref. [28], presents a comparative study of five deep learning (DL) methods to forecast the number of new cases and recovered cases; the promising potential of a deep learning model in forecasting COVID-19 is demonstrated. Similarly, in [4] a comparative study of DL and machine learning models for COVID-19 transmission forecasting was performed; experimental results showed that the best performance was archived by DL, especially the LSTM-CNN model (which is the combination of Long Short-Term Memory-LSTM and Convolutional Neural Networks-CNN).

Much work has been performed to forecast COVID-19 cases in different regions, and countries [15]. For example, Ref. [1] reported the forecast results of COVID-19 cases (obtained by Recurrent ANN-LSTM and Recurrent ANN-GRU models) throughout 60-day in ten countries (USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, United Kingdom, and Iran). Refs. [6, 12]) have followed the same direction, and they have shown experimental results of forecast covid-19 cases for multiple countries or populations. Also, specialized studies over particular countries have been developed; Ref. [23] proposes a recurrent and convolutional neural network model to forecast COVID-19 cases confirmed daily in 7, 14, and 21 days in India. Similarly, Ref. [21] studies statistical and artificial intelligence approaches to model and forecast the prevalence of this epidemic in Egypt.

In the Mexican context, some studies have been presented. In [22] cases of COVID-19 infection in Mexico are modeled and predicted through mathematical and computational models using only the confirmed cases provided by the daily technical report COVID-19 Mexico, from February 27th to May 8th, 2020. Ref. [11] uses ANN to predict the number of COVID-19 cases in Brazil and Mexico until December 12, 2020. In the same year, Ref. [18] presents an analysis of the ensemble of the neural network model with fuzzy response aggregation to predict COVID-19 confirmed cases of COVID-19 in 109 days ahead for Mexico (whole Country), which confirms other studies where ensemble method works better than monolithic classifiers, in this case on predicting the COVID-19 time series in Mexico. Another very interesting paper [8] compares traditional and powerful forecasting methods (vector autoregression and statistical curve-fitting methods) concerning DL techniques (in particular, the LSTM model) to identify the pandemic impact in Mexico in a period of 70 days (January 22 to March 31, 2020); it concludes that the best practice is to use LSTM over classical models.

In this paper, we present an empirical study of four popular ANN: Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and LSTM-CNN, to forecast COVID-19 cases in Mexico. These ANN have been reported in the state of art as the best models for this type of goal. We use recent COVID-19 data (from February 22, 2020, to April 4, 2022); therefore, we now have more information about COVID-19 than in previous works, suggesting an improvement

in the forecasting of COVID-19 cases. In this sense, this study intends to be conceived as a basis for comparing deep learning techniques in the context of similar problems, characterized by being highly sensitive to data variability in applications other than classical classification tasks. Results indicate that LSTM-CNN achieves the best qualitative performance, but the CNN model reports the best quantitative metrics.

2 Theoretical Framework

Feed-forward constitutes the most conventional ANN architecture. It is commonly based on at least three layers: input, output, and one hidden layer [16]. In the DL context, feed-forward DL-ANNs have two or more hidden layers in their architecture. This allows to reduce the number of nodes per layer and uses fewer parameters, but it leads to a more complex optimization problem [9]. However, this disadvantage is less restrictive than before due to the availability of more efficient frameworks, such as Apache-Spark or TensorFlow (which use novel technology like GPU or Cluster Computing). Another important DL model is recurrent neural networks, which are a type of network for sequential data processing, allowing to scale of very long and variable-length sequences [27]. In this type of network, a neuron is connected to the neurons of the next layer, to those of the previous layer, and to itself using weights (values that change in each time step). A summary of the DL models studied in this work is featured below.

2.1 MLP

MLP is a classical ANN with one input and one output layer, and at least one hidden layer, trained with a different set of features based on the previous layer's output (see Fig. 1). It is possible to select features in a first layer, and the output of this will be used in the training of the next layer [16]. The number of inputs and outputs of the problem to be solved is the factor that will determine the number of neurons in the input and output layers, and every neuron can feed into the next neuron of the next layer by repeating

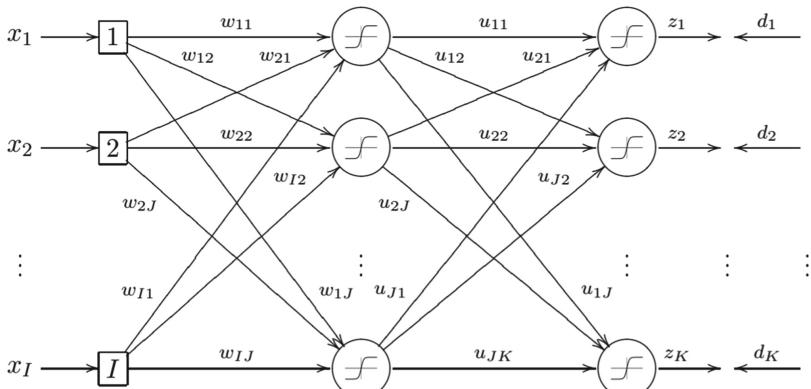


Fig. 1. Classical MLP architecture comprised by 3 layers, I input nodes, J hidden nodes and K output nodes.

the process from the input until the output layer [13]. An MLP structure can achieve significant performance in small models sizes, but when its size scales up, the model is affected by the over-fitting [31]. MLPs could approximate any continuous function and can solve not linearly separable problems.

2.2 CNN

CNN is a deep neural network architecture that combines the multilayer perceptron with a convolutional layer to build a map that has the function of extracting important features (see Fig. 2). Furthermore, it implements a pooling stage to reduce the dimensionality of the features and save the most informative features [3]. The main idea behind these models is that abstract features can be extracted by the convolutional layers and the pooling operation, where the convolutional kernels convolve local filters with sequential data without processing and produce non-variant local features, and the subsequent pooling layers will extract the essential features within fixed-length sliding windows [30], in other words, a CNN is a powerful extractor that applies convolution on multiple blocks to extract meaningful features [25]. CNN models have shown to be effective in problems related to modeling image data, summarization, and classification [14].

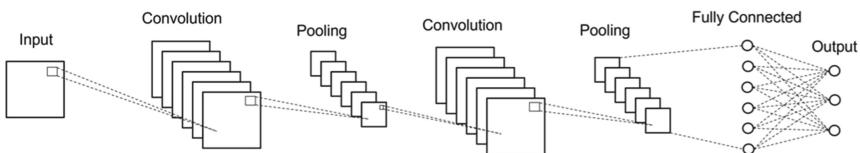


Fig. 2. A CNN architecture comprised by two convolutional layers and two pooling layers.

2.3 LSTM

LSTM network is a particular type of recurrent neural network that can learn in the long term to avoid dependency [19]. To achieve this, LSTM uses different cells to allow actions such as “forget” and “remember” [3]; in other words, LSTM units consist of elements such as an input gate, a forget gate, a memory cell, and an output gate [30] (see Fig. 3). It is important to say that LSTM was designed to prevent the backpropagating error from disappearing or exploding; likewise, forget gates were included to achieve long-term non-dependence, being able to control the use of state cell information [29]. These architectures were designed to work with data in constant times that occur between elements of a given sequence [2]. Due to its ability to capture long-range dependencies, this model has been successfully applied in many areas, such as speech recognition, handwriting recognition, image recognition, and natural language processing [29].

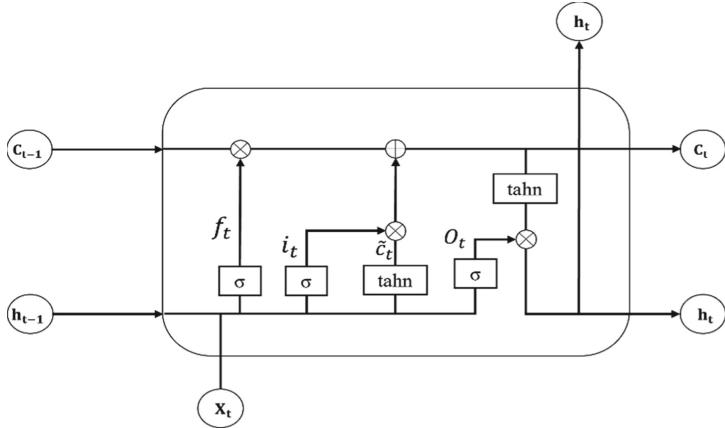


Fig. 3. Architecture of an LSTM unit with a forget gate (f_t), current input (X_t), memory cell (C_t) and output (h_t).

2.4 LSTM-CNN

In recent years, hybrid deep learning architectures have been applied to different tasks showing better results than the baseline models. A clear example of these hybrid frameworks are the LSTM and CNN architectures which have shown excellent performances in tasks such as time series classification, video recognition, and text classification due to their unique characteristics [24]. Combining these networks, the advantages of each one is merged to achieve more significant results [26]. The core idea behind the fusing of these models is that CNN can extract time-invariant elements, and LSTM can learn long-term dependency. Both LSTM and CNN receive the same data input, and then the results are concatenated to get the output [17] (see Fig. 4). Therefore, with this fusion, a better structure, and more complete spatial and temporal characteristics can be obtained, improving the results in the state of the art [7].

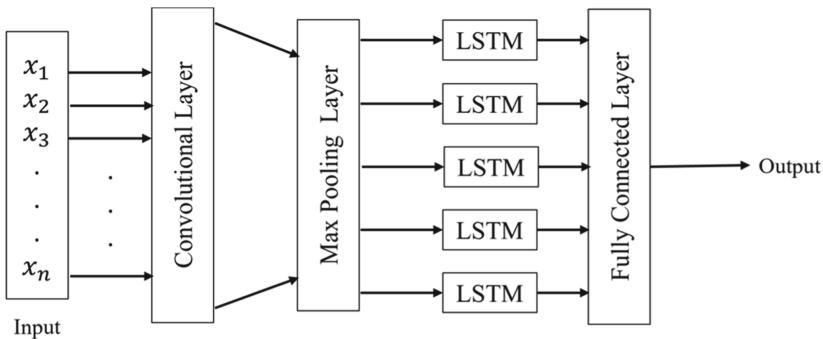


Fig. 4. Architecture of the LSTM-CNN hybrid model comprised by a convolutional layer and a LSTM block.

Traditionally, ANNs have been trained with the back-propagation algorithm (based on the stochastic gradient descent), and the weights are randomly initialized. However, in some late versions of DL-ANN, the hidden layers are pre-trained by an unsupervised algorithm, and the weights are optimized by the back-propagation algorithm [16] or methods based on the descending gradient. To overcome it, the classical Sigmoid activation function has been replaced (commonly) by other functions like Rectified Linear Unit (ReLU) $f(z) = \max(0, z)$, Exponential Linear Unit (ELU = $zif\{z \geq 0\} else \{\alpha * (e^z - 1)\}$), or softmax ($\varphi(s) = e^{s_i} / \sum_j^C e^{s_j}$), that is associated to the output layer), because typically they learn much faster in networks with many layers, allowing training of a DL-ANN without unsupervised pre-training [9].

The most common algorithms of descending gradient optimization are: a) Adagrad, which adapts the learning reason of the parameters, making more significant updates for less frequent parameters and smaller for the most frequent ones, b) Adadelta is an extension of Adagrad that seeks to reduce aggressiveness, monotonously decreasing the learning rate instead of accumulating all the previous descending gradients, restricting accumulation to a fixed size, and c) Adam, that calculates adaptations of the learning rate for each parameter and stores an exponentially decreasing average of past gradients. Other important algorithms are AdaMax, Nadam, and RMSprop [20].

3 Experimental Set Up

In this section, the experimental details to allow the proper replication of the results of this research and to support the conclusions, are described.

3.1 Dataset

The dataset was extracted from the GitHub repository (<https://github.com/CSSEGI/SandData/COVID-19>) from the Resource Center at the John Hopkins University of Medicine, which is updated daily at 9 am EST [5]. It is important to say that only the file of the confirmed cases of COVID-19 was downloaded, and then only the dates and their cases reported from February 22, 2020, to April 4, 2022, for Mexico were extracted. After data collection, the data was organized by week (considering seven dates to form a week), and then the average per week was obtained to work with a smooth curve. In addition, the dataset ($D = \{x_1, x_2, \dots, x_Q\}$) was split in two disjoint sets, one (TR) to train the model and other (TS) to test its generalized ability, i.e., $D = TR \cup TS$; $TR \cap TS = \emptyset$; and (TR) contains 70% of the samples of D , and TS contains the remaining 30%. Thus, each individual x_q data in D corresponds to the average number of active COVID-19 cases in Mexico in one specific week, i.e., the average number of cases every seven days.

3.2 Free Parameters Specification

The DL model to use is an MLP, and it was designed with an input layer of 5 nodes and 1 hidden layer with 3 nodes, both with the RELU activation and the output layer with linear activation. CNN topologies consist of two CNN layers, the first with 16 filters (or kernels) of dimension 4×4 and the second with 32 filters of 1×1 . A pooling layer of size

2 (using the MaxPooling method), and a dense layer of 33 nodes, were used. The LSTM model has a sequence length and an input dimension of 5 and 16 with RELU activation and recurrent sigmoid activation. In addition, the dropout method is applied after the input layer and before another LSTM layer with 64 nodes and a recurrent sigmoid layer; finally, a linear activation for the output layer is used. LSTM-CNN contains an LSTM layer with sequence length, an input dimension of 4 and 64, respectively, a CNN layer with 32 filters of 4×4 and a RELU activation, and finally, a dense layer with 1 hidden neuron and a linear activation function.

3.3 Performance of the DL Models

Mean Squared Error (*MSE*) and Mean Absolute Error (*MAE*) are metrics widely accepted to assess ANN in the prediction and approximation of functions [9]. In this work, we use both, *MSE* and *MAE* (Eq. 1) to test the effectiveness of studied DL models.

$$MSE = \frac{1}{N} \sum_{i=1}^N (t_i - z_i)^2; \quad MAE = \frac{1}{N} \sum_{i=1}^N ||t_i - z_i||; \quad (1)$$

where N is the total of samples, t_i is the desired output, and z_i is the actual or predicted output of the ANN for the sample i .

DL models were developed in Tensorflow 2.0 and Keras 2.3.1 frameworks. Adam was selected as the optimizer method with a batch size of 9 for CNN, 10 for LSTM, and 1 for MLP and LSTM-CNN. The stop criterion was 500 epochs.

4 Results and Discussion

The main results obtained in the experimental stage are presented and discussed in this section. Figures 5, 6, 7 and 8 show the graphs generated by the four ANN studied in this work (CNN, MLP, LSTM and LSTM-CNN, see Sect. 2). Axis x represents the analyzed time periods (in this work, it corresponds to an interval of a week, for more detail see Sect. 3.1), and axis y corresponds to forecast and real COVID-19 cases.

From a qualitative viewpoint, experimental results seem to note that the best performance corresponds to CNN (Fig. 5), where both: real and predicted values, are very similar. It matches with quantitative results, where MSE and MAE values of CNN are the smallest. However, considering that the dataset is small, this behavior may imply that overfitting could be occurring. In MLP (Fig. 6), this behavior is more evident; we observe smaller MSE and MAE in the training data (blue) than the obtained with test data (red); nevertheless, MLP follows the data trend from the qualitative viewpoint.

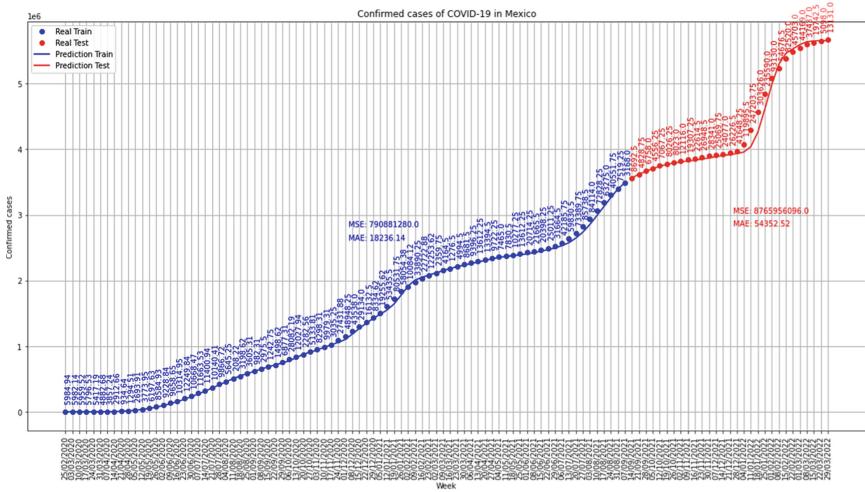


Fig. 5. Results obtained by the CNN model.

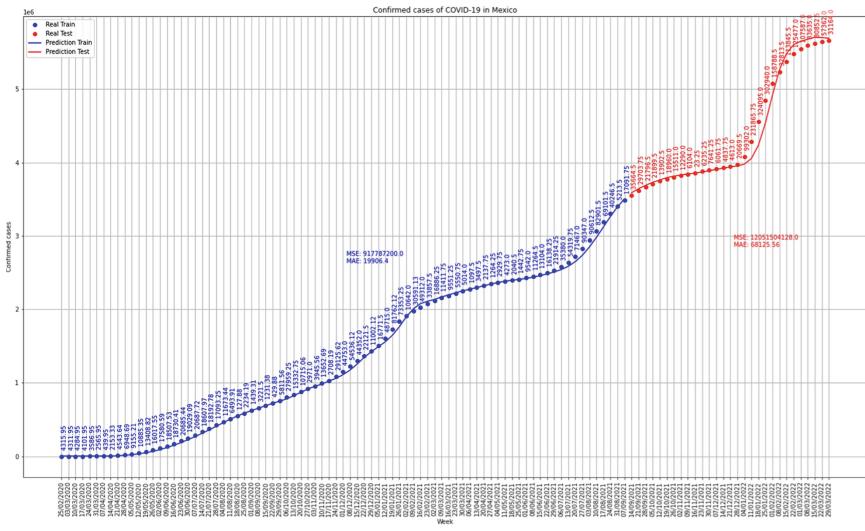


Fig. 6. Results obtained by the MLP model.

LSTM model exhibits a similar trend to MLP for the training dataset, but the behavior of LSTM in the test data does not match the actual data. It is reflected in the MAE and MSE values test, which are more significant than the MLP case. LSTM is characterized by long-term learning dependency. Thus, results presented in Fig. 7 do not notice that this feature of LSTM (by itself) is enough to approximate COVID-19 data with good performance.

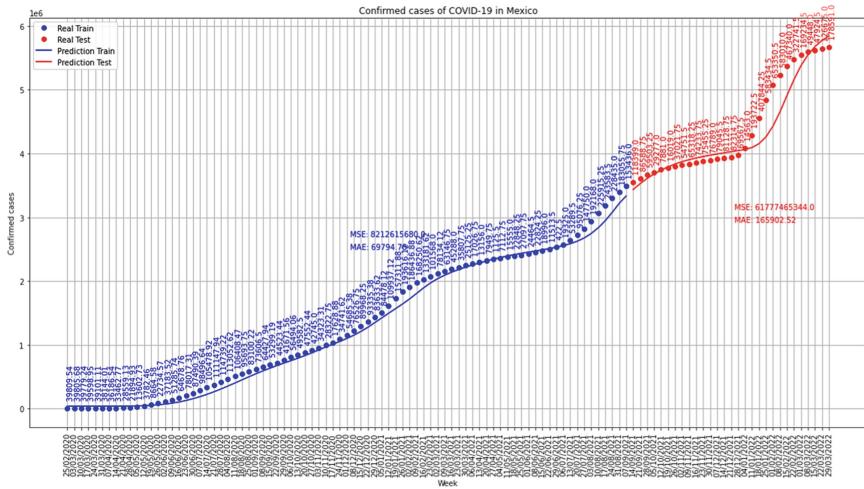


Fig. 7. Results obtained by the LSTM model.

LSTM-CNN shows a better performance from a qualitative viewpoint, as can be seen in Fig. 8 where, although the model does not entirely fit in the training and testing datasets, LSTM-CNN has the best generalization ability, which refers to the capability of the model to give an appropriate answer to unlearned questions. The generalization performance of LSTM-CNN could be explained by the combination of the long-term learning dependency of the LSTM model and the CNN's capacity to exploit the information extracted from the data.

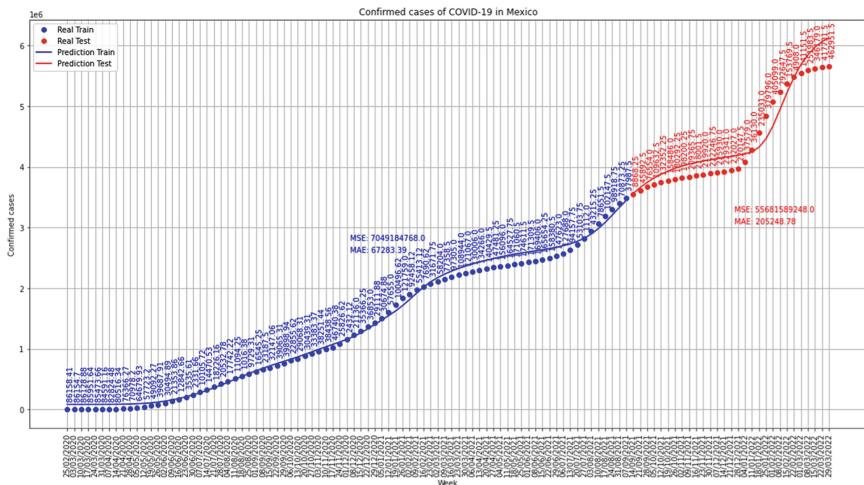


Fig. 8. Results obtained by the LSTM-CNN model.

Finally, in the recent state-of-art about the COVID-19 forecasting, the LSTM and LSTM-CNN appeared to report the better behavior, even overcoming the performance of the other models. However, the results presented in this work (which use more information about confirmed COVID-19 cases in Mexico than previous works) show that from a quantitative viewpoint, MLP and CNN models obtain smaller values of MSE and MAE than LSTM and LSTM-CNN. The best data fit performance is presented by the CNN model (Fig. 5), which would suggest that ability to long-term learning dependency data is not a critical aspect to forecast COVID-19 cases. Thus, from a quantitative viewpoint, results showed by MLP and CNN imply the best performance when the most relevant information is extracted from the data; MLP, however, it is considered that the performance of the CNN is attained due it has a better capability to transform the abstract feature space in the convolutional layers inherent to its architecture. This behavior could be explained by the fact that deep neural network potential is found in its hidden space, where it can abstract high-level patterns. In such hidden space, original data are transformed to another multi-dimensional space, in which the decision boundary could be identified with a higher degree of reliability [9, 16].

5 Conclusion

In this paper, we studied four ANN models: MLP, CNN, LSTM, and LSTM-CNN, to forecast COVID-19 cases. Experimental results analyzed from a qualitative viewpoint suggest that the LSTM-CNN model obtains the best performance. However, from the quantitative perspective, the CNN model overcomes the performance of MLP, LSTM, and LSTM-CNN. These results indicate that long-term learning dependency data is not critical to forecasting COVID-19 cases (see results of LSTM-CNN, but mainly of the LSTM model). Instead, the results exhibited by MLP and CNN imply that to obtain better performance is most important to extract relevant information from data features, which is a highlighted feature of MLP and CNN models.

The results presented in this work are exciting; nevertheless, future work is required to deep into this study and to develop a theoretical explanation for the experimental results due to the potential of the analyzed models to forecast COVID-19 cases in regions like Mexico or Latin America, which have been seriously affected by this pandemic.

References

1. ArunKumar, K., Kalaga, D.V., Kumar, C.M.S., Kawaji, M., Brenza, T.M.: Forecasting of COVID-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUS) and long short-term memory (LSTM) cells. *Chaos Solitons Fractals* **146**, 110861 (2021). <https://doi.org/10.1016/j.chaos.2021.110861>
2. Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient subtyping via time-aware LSTM networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 65–74. Association for Computing Machinery, New York (2017). <https://doi.org/10.1145/3097983.3097997>
3. Bengfort, B., Bilbro, R., Ojeda, T.: Applied Text Analysis with Python. O'Reilly Media, Inc., Sebastopol (2018)

4. Dairi, A., Harrou, F., Zeroual, A., Hittawe, M.M., Sun, Y.: Comparative study of machine learning methods for COVID-19 transmission forecasting. *J. Biomed. Inform.* **118**, 103791 (2021). <https://doi.org/10.1016/j.jbi.2021.103791>
5. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track covid-19 in real time. *Lancet Inf. Dis.* **20**(5), 533–534 (2020). [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1)
6. Fanelli, D., Piazza, F.: Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos Solitons Fractals* **134**, 109761 (2020). <https://doi.org/10.1016/j.chaos.2020.109761>
7. Gao, J., Gu, P., Ren, Q., Zhang, J., Song, X.: Abnormal gait recognition algorithm based on LSTM-CNN fusion network. *IEEE Access* **7**, 163180–163190 (2019). <https://doi.org/10.1109/ACCESS.2019.2950254>
8. Gomez-Cravioto, D.A., Diaz-Ramos, R.E., Cantu-Ortiz, F.J., Ceballos, H.G.: Data analysis and forecasting of the COVID-19 spread: a comparison of recurrent neural networks and time series models. *Cogn. Comput.*, 1–12 (2021). <https://doi.org/10.1007/s12559-021-09885-y>
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
10. Guo, M., Manzoni, A., Amendt, M., Conti, P., Hesthaven, J.S.: Multi-fidelity regression using artificial neural networks: efficient approximation of parameter-dependent output quantities. *Comput. Methods Appl. Mech. Eng.* **389**, 114378 (2022). <https://doi.org/10.1016/j.cma.2021.114378>
11. Hamadneh, N.N., Tahir, M., Khan, W.A.: Using artificial neural network with prey predator algorithm for prediction of the COVID-19: the case of Brazil and Mexico. *Mathematics* **9**(2) (2021). <https://doi.org/10.3390/math9020180>
12. Hamdy, M., Zain, Z.M., Alturki, N.M.: COVID-19 pandemic forecasting using CNN-LSTM: a hybrid approach. *J. Control Sci. Eng.* **2021**, 8785636 (2021). <https://doi.org/10.1155/2021/8785636>
13. Kahani, M., Ahmadi, M.H., Tatar, A., Sadeghzadeh, M.: Development of multilayer perceptron artificial neural network (MLP-ANN) and least square support vector machine (LSSVM) models to predict Nusselt number and pressure drop of TiO₂/water nanofluid flows through non-straight pathways. *Numer. Heat Transf. Part A Appl.* **74**(4), 1190–1206 (2018). <https://doi.org/10.1080/10407782.2018.1523597>
14. Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S.: Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.* **173**, 24–49 (2021). <https://doi.org/10.1016/j.isprsjprs.2020.12.010>
15. Kuvvetli, Y., Deveci, M., Paksoy, T., Garg, H.: A predictive analytics model for COVID-19 pandemic using artificial neural networks. *Decis. Anal. J.* **1**, 100007 (2021). <https://doi.org/10.1016/j.dajour.2021.100007>
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015). <https://doi.org/10.1038/nature14539>
17. Li, P., Abdel-Aty, M., Yuan, J.: Real-time crash risk prediction on arterials based on LSTM-CNN. *Accid. Anal. Prev.* **135**, 105371 (2020). <https://doi.org/10.1016/j.aap.2019.105371>
18. Melin, P., Monica, J.C., Sanchez, D., Castillo, O.: Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. *Healthcare* **8**(2) (2020). <https://doi.org/10.3390/healthcare8020181>
19. Muzaffar, S., Afshari, A.: Short-term load forecasts using LSTM networks. *Energy Procedia* **158**, 2922–2927 (2019). <https://doi.org/10.1016/j.egypro.2019.01.952>. Innovative Solutions for Energy Transitions
20. Ruder, S.: An overview of gradient descent optimization algorithms. *CoRR* abs/1609.04747 (2016). <http://arxiv.org/abs/1609.04747>

21. Saba, A.I., Elsheikh, A.H.: Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Saf. Environ. Prot.* **141**, 1–8 (2020). <https://doi.org/10.1016/j.psep.2020.05.029>
22. Torrealba-Rodríguez, O., Conde-Gutiérrez, R., Hernández-Javier, A.: Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. *Chaos Solitons Fractals* **138**, 109946 (2020). <https://doi.org/10.1016/j.chaos.2020.109946>
23. Verma, H., Mandal, S., Gupta, A.: Temporal deep learning architecture for prediction of COVID-19 cases in India (2021)
24. Vo, Q.H., Nguyen, H.T., Le, B., Nguyen, M.L.: Multi-channel LSTM-CNN model for Vietnamese sentiment analysis. In: 2017 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 24–29 (2017). <https://doi.org/10.1109/KSE.2017.8119429>
25. Wu, J.L., He, Y., Yu, L.C., Lai, K.R.: Identifying emotion labels from psychiatric social texts using a bi-directional LSTM-CNN model. *IEEE Access* **8**, 66638–66646 (2020). <https://doi.org/10.1109/ACCESS.2020.2985228>
26. Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., Li, F.: Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Syst. Appl.* **169**, 114513 (2021). <https://doi.org/10.1016/j.eswa.2020.114513>
27. Yu, Y., Si, X., Hu, C., Zhang, J.: A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**(7), 1235–1270 (2019). https://doi.org/10.1162/neco_a_01199
28. Zeroual, A., Harrou, F., Dairi, A., Sun, Y.: Deep learning methods for forecasting COVID-19 time-series data: a comparative study. *Chaos Solitons Fractals* **140**, 110121 (2020). <https://doi.org/10.1016/j.chaos.2020.110121>
29. Zhao, R., Wang, J., Yan, R., Mao, K.: Machine health monitoring with LSTM networks. In: 2016 10th International Conference on Sensing Technology (ICST), pp. 1–6 (2016). <https://doi.org/10.1109/ICST.2016.7796266>
30. Zhao, R., Yan, R., Wang, J., Mao, K.: Learning to monitor machine health with convolutional bi-directional LSTM networks. *Sensors* **17**(2) (2017). <https://doi.org/10.3390/s17020273>
31. Zhao, Y., Wang, G., Tang, C., Luo, C., Zeng, W., Zha, Z.: A battle of network structures: an empirical study of CNN, transformer, and MLP. CoRR abs/2108.13002 (2021). <http://arxiv.org/abs/2108.13002>



Multi-view Robustness-Enhanced Weakly Supervised Semantic Segmentation

Yu Sang^(✉), Shi Li, and Yanfei Peng

School of Electronic and Information Engineering, Liaoning Technical University,
Huludao 125105, China
sangyu2008bj@sina.com

Abstract. Semantic segmentation is the basic work in computer vision; it has been shown to achieve adequate performance in the past few years. However, owing to the inherent logical obstacles in the classification architecture, it lacks the ability to understand the long-distance dependence in the image. To address this issue, we propose a new architecture to allow the model to more expansively mine the available information for classification and segmentation tasks in a weakly supervised manner. Firstly, we raise a masking-based data enhancement approach, where images are randomly masked based on scale, forcing the model to observe other parts of the object. Secondly, a long-range correlation matrix is introduced from the image itself to make the class activation mapping (CAM) a more complete coverage on foreground objects. Finally, the experimental results on the PASCAL VOC 2012 dataset show that our method can better exploit the salient parts and non-salient regions of foreground objects in weakly labeled images comparing with other methods. On the test set, our approach achieves mIoUs of 60.9% using ResNet-based segmentation models, outperforming other methods for the weakly supervised semantic segmentation (WSSS) task.

Keywords: Weakly supervised semantic segmentation (WSSS) · Class activation mapping (CAM) · Data enhancement · Self-attention mechanism

1 Introduction

The pipeline of semantic segmentation is to separate an image into regions, predict all pixels in each region, and assign semantic categories to them [1–3]. As a fundamental work in the field of computer vision, semantic segmentation-based algorithms have been widely used in practice such as intelligent surveillance, remote sensing images and automatic vehicle control. Currently, most of the researches in the view of deep convolutional neural networks are fully supervised methods and have achieved excellent results, while the prerequisite for obtaining the desired accuracy is having a dataset with high quality and large amount of data. Unfortunately, manual labeling is costly and can only be obtained by spending a lot of time and effort. Therefore, weakly supervised models have become a hot topic of interest for researchers, which require only simple annotation to accomplish the effect of full supervision, such as scribble-based annotation

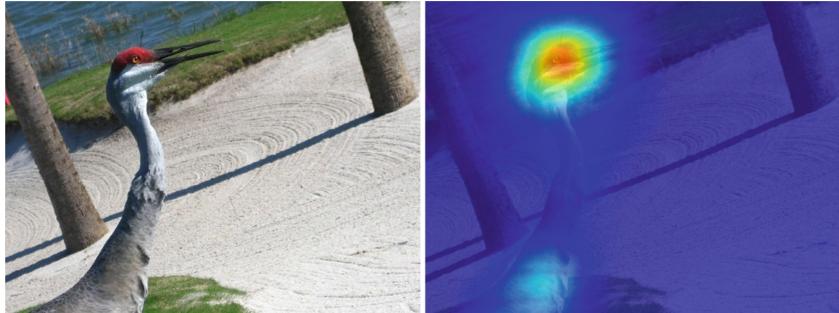


Fig. 1. The left image shows a picture with the semantic category bird, and the right image shows the class activation mapping (CAM) obtained by the classification network.

[4,5], point-level annotation, bounding box annotation [6, 7], and image-level annotation [8–10].

Existing excellent work on weakly supervised semantic segmentation (WSSS) based on image-level labeling usually follows the pipeline of generating pseudo-masks for the target object via class activation mapping (CAM) [11] and the generated pseudo-masks are then utilized to further train the desired segmentation model using some method in a fully supervised manner [1]. Unfortunately, CAM tends to identify small parts of objects rather than entire regions, resulting in only sparse and incomplete seeds (also known as pseudo-annotations). As shown in Fig. 1, an image with the semantic category bird is passed into the classification network for feature extraction, and its corresponding pixel-level pseudo-labeling is generated using the CAM method. However, there is an inherent problem in the transformation from classification network to segmentation network, CAM does not extract the whole object contour of the bird completely, but will focus the response on the most significant region of the object (e.g., the head of the bird), which is not sufficient for subsequent semantic segmentation. Therefore, to alleviate this phenomenon, quite a few excellent methods have been proposed to recover dense and reliable seeds. In AE-PSL [12], Wei et al. modify the structure of the network so that the model erases the high response region of the object in the previous round of sub-training and in the next training phase, the model can only complete the logical operations in the remaining parts of the picture to mine the information of the object, so as to obtain rich pseudo masks. Kim et al. [13] uses the discriminative region suppression module to reduce the activation value of high recognition parts and narrow the gap between them and adjacent non-significant regions. Thus, the coverage area of significant areas in the whole region is expanded in a disguised manner. In BEnet [14], the model uses the initially obtained localization maps (also CAMs) to generate boundary regression results in an additional branch of the boundary exploration network that explicitly obtains the object boundaries and later uses them to impose more favorable constraints on the propagation of CAMs. SEAM [15] noticed that in weakly supervised tasks, data enhancement methods performed on images is not effective owing to the lack of pixel-level supervision corresponding to the input image. Therefore, Wang et al. [15] uses a Siamese network into which the original image is passed simultaneously with

its enhanced version and introduces a self-supervised equivariant loss to strengthen the supervision of the model.

Since the supervision of weakly supervised classification network is image level, they only provide simple information, such as whether an object presence or not, rather than the full contour of the object, this causes difficulties for the classification network to capture the full contextual information of the foreground objects in certain complex scenes. In this paper, we comply with the previous approach and construct a weakly supervised segmentation network, whose inputs are image-level annotations. The random masking taken on the input image allows the classification model to focus on the full view of the object rather than a small number of salient regions. A contextual attention module is introduced to compute and obtain the self-correlation matrix between pixels at any two points in the input image and guide the CAM to respond to pixels with the same or similar semantics. For the corresponding CAMs obtained for multiple inputs. We propose object consistency loss to provide self-supervised constraints. The notable contributions of our paper are briefly described as follows:

- We introduce an erased-based data enhancement approach that allows the network model to over-expand the region of interest during classification task. To make it applicable to weakly supervised data, a multiple-input approach is adopted.
- In order to further expand the region of CAM, a contextual attention module is introduced to measure the correlation and similarity of inter-pixel to correct for low activation values in non-significant regions. To hold back the model from crashing due to the introduction of too much noise, an object consistency loss is proposed to constrain the model at the pixel level according to the reliable CAM.
- Experiments on the PASCAL VOC 2012 dataset show that the performance of the image segmentation task has been improved. Excellent results are achieved on public data sets compared to other methods of segmentation algorithms.

The rest of this paper is organized as follows. In Sect. 2, we present our literature review on WSSS and other related tasks. Section 3 describes our proposed method in detail. Section 4 then details the experimental setup used to evaluate our approach and shows our experimental results. Finally, the main conclusions drawn from the work are summarized.

2 Related Work

Previous approaches to images semantic segmentation were largely based on simple features of the object, such as the shape, color, and grayscale, using edge-based, threshold-based, and region-based methods. These approaches can use only the lower level of visual information of the image itself and are too dependent on the segmented object, which is difficult to use in complex realistic scenarios. Long et al. [2] first proposed the model that relies on convolutional operations to perform segmentation and the segmentation accuracy exceeded that of contemporaneous algorithms. In the 2015 MICCAI Medical Image Segmentation Challenge, Ronneberger et al. [16] built U-Net to apply an encoder-decoder network to the segmentation task, where the image is compressed by an

encoder and later restored to its original resolution by a deconvolution operation. U-Net provides rich shallow information to those features that after decoding process through skip connections to solve the problem of spatial scale information loss caused by down-sampling. On this basis, Wang et al. [17] addressed the phenomenon that a large amount of effective information would be lost in the process of constant up-sampling and down-sampling by adopting the high-resolution representation strategy, which accompanies the deepening of the network with multiple branches of resolution in parallel to achieve the purpose of enhancing semantic information and obtaining accurate location information. At this point, the mainstream convolutional neural network model for building semantic segmentation tasks is basically established.

Over the past few years, semantic segmentation has been relatively well developed, and the fully supervised semantic segmentation algorithm can achieve 90.5% accuracy on the PASCAL VOC 2012 dataset. However, the cost of annotation the dataset required for fully supervised semantic segmentation is expensive, and the easy availability of weakly supervised labels has led an increasing number of researchers to devote into WSSS based algorithms. WSSS algorithms only require weakly supervised labels in training, which greatly reduces the expensive cost for training networks. The characteristic of weakly supervised learning method is that the label data it uses can only represent the existence of a semantic object, but does not provide specific information about the location or boundary of the object. These labels are weaker than pixel level labels. Naturally, when a large amount of data is required, it can be obtained with relatively low human and material resources. However the information obtained from image-level labeling alone is not sufficient to distinguish complete objects. Therefore, instead of pursuing to adjust the model so that it generates a perfect pseudo-labeling, Fan et al. [18] used different scale inputs and the fusion of multiple methods to obtain separate pseudo-seeds, and it is highly likely that there are complementary parts in multiple different seeds, which provides clues for inferencing and makes the foreground objects in the pseudo-label more complete. Fan's method is simple and effective; however multiple rough pseudo-labels will enhance the noise together. Jiang et al. [19] observed that the regions with high CAM activation shift continuously as the training process proceeds, so online attention accumulation was used to obtain a larger foreground localization map by overlaying the CAMs obtained in each training round.

Due to the limited supervision of weak labels, some approaches use additional data to enhance the accuracy of the model. For example, the object saliency map is introduced as a priori knowledge used to guide the network. Lee et al. [20] proposed a saliency loss to constrain the mismatch between the pseudo-labeled boundaries generated by the model and the true values, reinforcing the algorithm to accurately learn target boundaries. Wu et al. [21] constructed multiple classification functions in a deep neural network to explicitly combine the results of CAM with the classification task, which generates category-independent masks by using salient activation layers, and subsequently aggregates contextual information within and between images through a multi-headed attention aggregation mechanism. Fan et al. [22] considered that the segmentation task needs to rely on CAM to locate the target, while in image salient segmentation, the object that may be the foreground should be separated from the background, the two

tasks can be mutually reinforcing. Therefore, the joint multi-task framework is proposed to accomplish the WSSS task as well as the saliency detection task simultaneously.

When using image-level annotation as supervision to establish semantic segmentation model, the core idea of the above methods is the same, i.e., generating pixel-level high-quality pseudo-masks in the first step and use the pseudo-masks as fully supervised annotations in the second step to train the segmentation models. However, the two-step approach comes at the cost of model complexity and multi-stage training, so some scholars construct end-to-end frameworks to carry out semantic segmentation and enhance segmentation accuracy by other constrained optimization strategies. For example, Araslanov et al. [23] proposed a single-stage WSSS model that does not rely on additional saliency estimation, instead using normalized global weighted pooling to compute classification results, pixel-adaptive mask refinement to provide self-supervised constraints, and a stochastic gate to mix deep and shallow features in the training phase. Zhang et al. [24] modified the traditional end-to-end semantic segmentation network to have two branches: one branch completes the classification task and generates pixel-level pseudo labels through CAM, and the other branch then uses it to complete the segmentation. In contrast to the previous two-step method that tends to mine dense and complete object regions, Zhang's approach identifies reliable regions based on confidence and further prunes them by conditional random fields, and then uses it as a supervision means of parallel semantic segmentation branches.

3 The Proposed Approach

In this section, we first describe how we reinforce the classification model to excavate information during training phase. Then give an account of our approach to segmentation using multiple noisy images. Finally, we discuss our contextual attention module in detail.

3.1 Overview



Fig. 2. The left, middle and right images are the images to be segmented, its ground truth, and the result obtained by CAM, respectively. CAM shows that the response map will always be located in the most easily identifiable part of the object.

In Fig. 2, it can be seen that for an image with two objects, the model focus on the human’s head and the horse’s head respectively. In other words, only these parts need to be known for the model to obtain the optimal solution for the classification task. This phenomenon is known as information bottleneck: The input information is compressed as much as possible when passing through the deep neural network, while keeping as much task-relevant information as possible, which facilitates obtaining the best representation for classification tasks. Information bottlenecks prevent classification logic operations from considering non-significant information about the target object, resulting in CAMs that focus only on distinct regions of the target object. The key to matching both partial and complete features is to allow the network to reason about foreground objects using sufficient information.

3.2 Network Framework

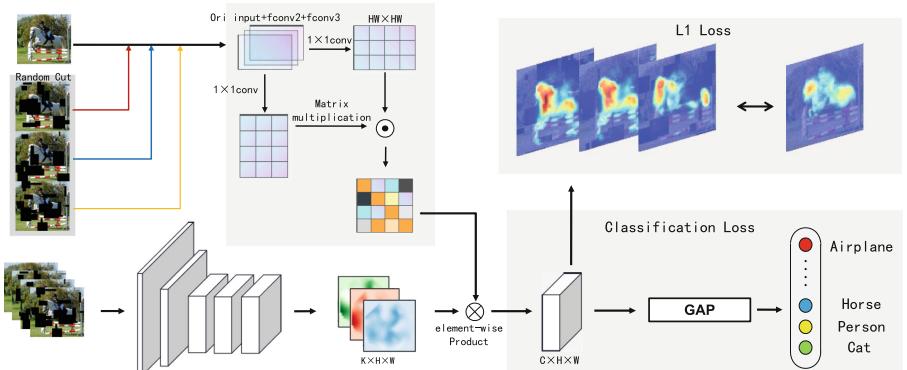


Fig. 3. The framework of our network. The pipeline for WSSS designed to complete the classification regression. An original image is subjected to the training process along with its three noisy versions. And they compute their own correlation matrix separately. Finally, the different class activation maps obtained will be weighted averages, and then the gap between them will be calculated with the initial CAM.

The overall framework is shown in Fig. 3. Firstly, a multi-class classification network is trained using weakly labeled data $\{I, L\}$, where I denote the input image and L is its corresponding image-level annotation. The image is computed by stacked convolutional layers and a feature map is obtained. Let K be the number of channels of the feature map and H, W be its spatial resolution. To generate CAM, by taking into account the dependency of object semantics, the contextual attention mechanism is used to improve the expressiveness of the network by appropriately tuning the features. The integrated feature map will adjust the channel to C by a classification layer. Finally, global average pooling is applied to compress the information and obtain the final prediction vector $l \in C$. To enhance the model’s ability to extract features, Inspired by Fan et al. [18], the original image is erased according to a scale-based erasure method to obtain several different noisy versions I_r . The random position coordinates $\{x, y\}$ and the masking

size $\{h, w\}$ are selected to control the masking area by applying the scale λ . Since I_r still corresponds to the semantic L , the model extracts features in the remaining part of the image and predicts its result. In consideration of that the masking process selects the coordinates of the masking locations randomly, for some images it may appear that a large portion of the foreground object is masked and thus the category of the image does not match its category. To prevent the introduction of model classification errors and improve the stability of training, multiple noisy inputs are used to compensate for the errors caused by the introduction of misinformation (Fig. 4).

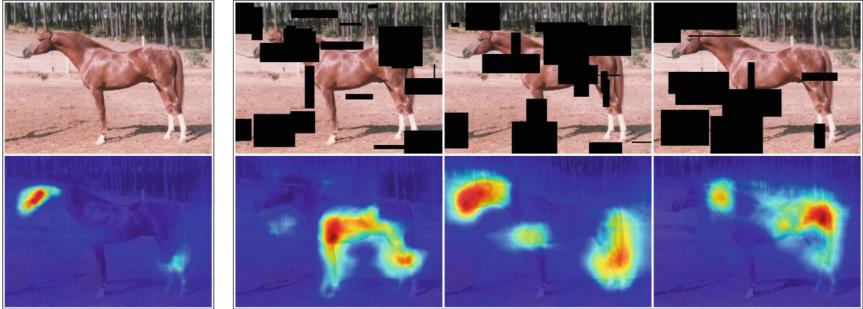


Fig. 4. The size of the masked area is controlled using a scale, and the coordinates are generated by random. Prominent parts of the object may be blocked. After cropping, the model will be forced to focus on other areas.

Owing to the perceptual field of the convolution kernel is limited in extracting features, many down-sampling layers need to be accumulated to associate different parts of the regions in the whole image, and successive pooling operations reduce the resolution of features while increasing the perceptual field. So the classification network is weak in capturing complex contextual information. Therefore we hope to make full use of the available information to capture more appropriate contextual information using the input image and shallow features as a guidance to mine the non-significant regions. The formula is as follows, whose inputs are the last layer feature map Y of the network and the tensor X obtained after stitching the original image with the shallow features $fconv2$ and $fconv3$.

$$Y_{\text{refine}} = \frac{Y\xi(X)}{\sum_{i=1}^{HW} \sum_{j=1}^{HW} \xi(X)_{ij}} \quad (1)$$

$$\xi(X) = \text{ReLU} \left(\frac{f(X_i)^T f(X_j)}{\|f(X_i)\| \cdot \|f(X_j)\|} \right) \quad (2)$$

The Formula (2) measures inter-pixel feature similarity by cosine distance. A convolution operation $f(x)$ is added, which is aiming at adjust the tensor dimension and transpose it afterwards to perform matrix multiplication. The similarity is deflated to the $[-1, 1]$ interval using the normalization operation to reduce the complexity of the subsequent computation. Similarly, we use the ReLU activation function and $L1$ normalization to filter out negative values, because we hope that the similarity matrix can

judge which pixels are highly correlated rather than highly independent. The obtained final self-attention matrix is fully connected, which represents the degree of correlation between any two pixels i and j in the image. The final result Y_{refine} is formed by concatenating its own backup and the result of multiplying it with the similarity matrix according to the channel dimension. Thus, modeling based on object contextual semantic dependencies helps to improve the integrity of the feature map.

3.3 Loss Function

In optimizing our proposed network, comply with other multi category classification tasks; we apply multi label soft margin loss to calculate the gap between the predicted results and the ground truth. It can be written as follow:

$$l_{cls} = - \sum_{c=1}^C y_c \log \sigma(p_c) + (1 - y_c) \log [1 - \sigma(p_c)] \quad (3)$$

$$L_{cls} = \frac{1}{N} (l_{cls}(x^0, y_c) + (N - 1)l_{cls}(\bar{x}, y_c)) \quad (4)$$

Formally, it is used for multi-category cross-entropy loss. Where x_0 and \bar{x} are the original pure image and the processed image respectively, P_c is the one-hot prediction value of the model for objects belonging to category C after the CAM obtained by Global Averaging Pooling, C is the foreground object category with a total of 20 classes, and y_c is the binary true value of the image corresponding to category C . The classification loss allows the model to perform parameter updates, which ultimately improves the accuracy of the model. The images after randomly masked, there will be cases where the foreground objects are masked over a large area or even disappear altogether, so we superimpose the results obtained from multiple masks and then calculate them with the CAM of the original image. We then construct the $L1$ paradigm to calculate the differences between the final CAMs obtained from different inputs.

$$L_{oc} = \|cam^{ori} - \sum cam^{hide}\|_1 \quad (5)$$

In summary, the network is jointly fine-tuning through the two items mentioned above, the equation is shown as:

$$L_{total} = L_{cls} + L_{oc} \quad (6)$$

4 Experiments

4.1 Dataset and Evaluation Metric

The PASCAL VOC 2012 dataset is used to verify our suspicions, which has 21 semantic classes, of which 20 are foreground objects and 1 is a background class. In our work, we follow the official practice of dividing the dataset into three parts: 1464 for training, 1449 for validation and 1456 for testing. In line with the approach of other methods, we

also add additional annotations from the semantic boundary dataset to build an enhanced training set, resulting in a training set of 10,582 images. During the training phase, only image-level annotations were used as a supervised approach.

The standard class-wise means intersection over union (mIoU) [15, 19] is adopted as evaluation metric for all the experiments.

4.2 Implementation Details

The images are enhanced before being fed into the model, including random scale scaling, random color shifting, horizontal flipping and finally cropping to a resolution of 448×448 . Our model is built using the Pytorch framework, using Resnet38 as the backbone network, and implemented using a single RTX 3090 GPU. A SGD optimizer was performed for optimization, which was initialized with a learning rate of 0.01 and allowed to decay according to a weight of 0.0005 per round as the model was trained. To further improve performance, the loss values are ranked using online hard example mining in each small batch, and the top 20% are taken as hard samples for training.

4.3 Ablation Study

To investigate the impact of each of the proposed components, we conducted comprehensive experiments in different settings on the PASCAL VOC 2012 while keeping the ResNet-based model constant as a guideline. We construct to adjust the ratio between the shadows and the original image, add black blocks randomly in the foreground to make it as the background, and observe whether we benefit from it by comparing the different results to the segmentation model. The length and width of each unique small shadow are calculated from the resolution of the original image, and they are distributed in a max-min interval. They can overlap each other, but the total shading area is to be constrained. We constructed experiments to verify the effects produced by different sizes, first letting the scales be arranged from 0.1 to 0.3. On this basis, add experiments to explore the effect of single to multiple images (Table 1).

Table 1. Experiments on various ratios

Masking ratios	mIoU on basic models (%)	
	One original image and one masked input	Multiple inputs with original image
0	47.82%	—
0.1	32.78%	46.61%
0.2	33.62%	47.76%
0.3	32.02%	48.01%

When only one image with noise is participating in the training at the same time as the original image, the model only tends to be collapsed and the result obtained is more like that the model is guessing what the pixel belong to. The inference accuracy of the model decreases as the ratio increases. The possible reason for this is that the foreground object disappears after processing and the current semantics of the image no longer corresponds to the real ground truth. The label mismatch causes the model to learn irrelevant feature information. And when we add multiple images for training again, the situation change for better. This is because when multiple inputs are used for the same image, for each input, the likelihood of the above situation occurring simultaneously is greatly reduced. In our experiments, we use a total of four input images, one pure image, and three additional processed versions.

Table 2. The ablation study for each component of our approach

Methods	mIoU (%)
Backbone	47.82
+ Multiple masked input	48.01
+ Contextual attention module	56.66
+ Object consistency loss	47.04
+ Multiple masked input + contextual attention Module + object consistency loss	57.53

To make the proposed method more convincing, we report the influence of each module in Table 2 with mIoU as the evaluation metric. The experimental results are used to analyze the effect of these three different modules on the segmentation accuracy. The training set and the number of iterative rounds used for the ablation experiments are consistent with the baseline. Comparing to CAM baseline which achieves 47.82%, the proposed multiple inputs improves mIoU by 0.19%. Applying Contextual attention module, the localization maps performance is significantly improved from 47.82% to 56.66%. Additionally, Object consistency loss brings 0.8% reduction, It shows simply single module cannot bring significant improvement. The method in this paper is the effect produced by adding these four modules simultaneously.

4.4 Comparison with Other SOTA Methods

To further explore the effectiveness of our proposed method, we conduct extensive comparisons with recently developed methods dedicated to learning from weak labels, such as AE-PSL, Affinity net, and etc. Table 3 exhibits the results compared to the previous method. For the sake of fairness, the backbone networks they adopted and results are those given in the original article. Table 4 lists the parameter quantities

of models. The subjective segmentation results are shown in Fig. 5. And finally the comparison of our method with the fully supervised method is placed in Fig. 6.

Table 3. Comparison with other SOTA methods

Methods	Backbones	Val	Test
CCNN [9]	VGG16	35.3	35.6
EM-Adapt [25]	VGG16	38.2	39.6
MIL [10]	OverFeat	42.0	40.6
SEC [8]	VGG16	50.7	51.7
STC [26]	VGG16	49.8	51.2
AdvErasing [12]	VGG16	55.0	55.7
Affinity net [28]	ResNet38	58.4	60.5
GAIN [20]	ResNet38	55.3	56.8
MCOF [27]	ResNet38	56.2	57.6
Deeplab* [1]	ResNet50	76.9	–
Ours	ResNet38	58.4	60.9

* means it is a fully supervised algorithm

Table 4. Complexity and Flops on several tasks

Architecture	Params	Flops
Affinity net [28]	96.7M	400.23G
MCOF [27]	79.4M	524.6G
Deeplab* [1]	57.8M	395.2G
Ours	100.5M	640.5G

* means it is a fully supervised algorithm

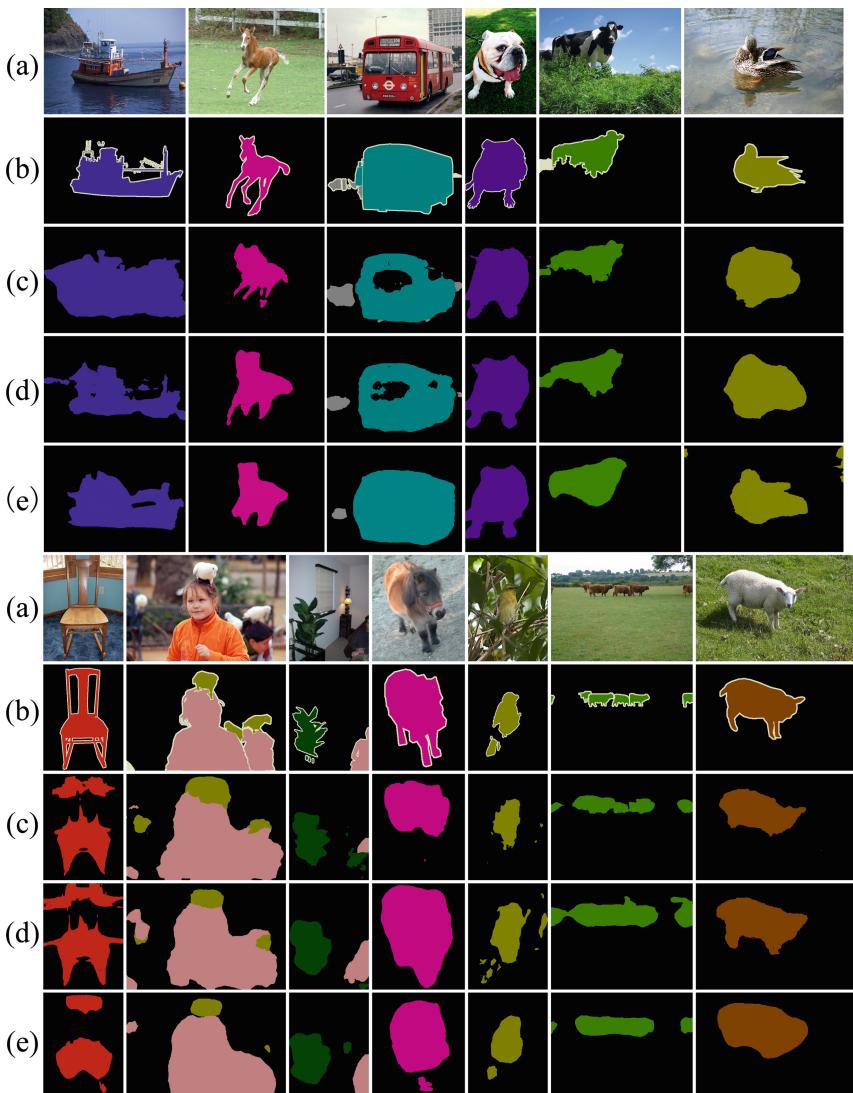


Fig. 5. Qualitative segmentation results on PASCAL VOC 2012 val set. (a) Original images. (b) Ground truth. (c) Segmentation results predicted by Affinity net. (d) Segmentation results predicted by MCOF. (e) Results of Ours-ResNet38. Our method better captures larger object areas and less prone to miss objects. However, the object boundaries in our results are somewhat larger than those obtained by other methods.

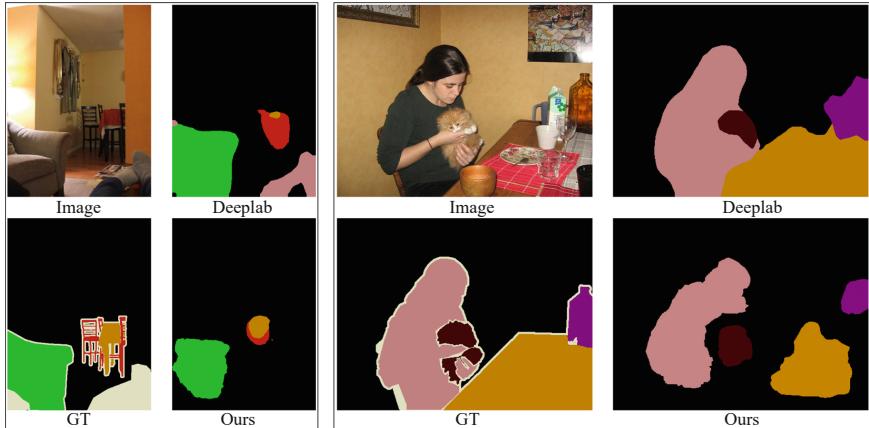


Fig. 6. Visual comparison with fully supervised algorithm

5 Conclusions

We propose a data enhancement approach for training with noisy data, which is a learning approach for training weakly supervised methods. To be brief, the input image is partially erased. This reduces the local optimization caused by training CNNs with the salient parts in the image. In addition, the relevant information of pixels in the image is obtained from the nature of the image itself, which makes CAM more accurate in specifying objects. Our study finalizes the semantic segmentation task by not using expensive annotated data for classification.

Acknowledgements. This work was supported in part by the Basic Scientific Research Project of Liaoning Provincial Department of Education under Grant No. LJKQZ2021152; in part by the National Science Foundation of China (NSFC) under Grant No. 61602226; in part by the PhD Startup Foundation of Liaoning Technical University of China under Grant No. 18-1021; in part by the Basic Scientific Research Project of Colleges and Universities in Liaoning Province under Grant No. LJKZ0358.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., 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 (2018)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2015)
3. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6230–6239 (2016)
4. Dai, J., He, K., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1635–1643 (2015)

5. Khoreva, A., Benenson, R., Hosang, J., et al.: Simple does it: weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1665–1674 (2016)
6. Di, L., Dai, J., Jia, J., et al.: ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3159–3167 (2016)
7. Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2953–2961 (2017)
8. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 695–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_42
9. Deepak, P., Philipp, K., Trevor, D., et al.: Constrained convolutional neural networks for weakly supervised segmentation. In: IEEE International Conference on Computer Vision, pp. 1796–1804 (2015)
10. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1713–1721 (2015)
11. Zhou, B., Khosla, A., Lapedriza, A., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
12. Wei, Y., Feng, J., Liang, X., et al.: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6488–6496 (2017)
13. Kim, B., Kim, S.: Discriminative region suppression for weakly-supervised semantic segmentation. In: AAAI 2021 (2021)
14. Chen, L., Wu, W., Fu, C., Han, X., Zhang, Y.: Weakly supervised semantic segmentation with boundary exploration. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12371, pp. 347–362. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58574-7_21
15. Wang, Y., Zhang, J., Kan, M., et al.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12272–12281 (2020)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Sun, K., Xiao, B., Liu, D., et al.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5686–5696 (2019)
18. Fan, J., Zhang, Z., Tan, T.: Employing multi-estimations for weakly-supervised semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 332–348. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_20
19. Jiang, P.T., Hou, Q., Cao, Y., et al.: Integral object mining via online attention accumulation. In: International Conference on Computer Vision (2019)
20. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9215–9223 (2018)

21. Wu, T., Huang, J., Gao, G., Wei, X., et al.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 16760–16769 (2021)
22. Fan, R., Hou, Q., Cheng, M.-M., Yu, G., Martin, R.R., Hu, S.-M.: Associating inter-image salient instances for weakly supervised semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 371–388. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_23
23. Nikita, A., Stefan, R.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4253–4262 (2020)
24. Zhang, B., Xiao, J., Wei, Y., et al.: Reliability does matter: an end-to-end weakly supervised semantic segmentation approach. [arXiv:1911.08039](https://arxiv.org/abs/1911.08039) (2019)
25. Papandreou, G., Chen, L.C., Murphy, K.P., et al.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4981–4990 (2016)
26. Wei, Y., Liang, X., Chen, Y., et al.: STC: a simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(11), 2314–2320 (2017)
27. Wang, X., You, S., Li, X., et al.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1354–1362 (2018)
28. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4981–4990 (2018)



Rolling Bearing Fault Diagnosis Based on Graph Convolution Neural Network

Yin Zhang¹ and Hui Li^{1,2(✉)}

¹ School of Mechanical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
Huili68@163.com

² Tianjin Key Laboratory of Intelligent Robot Technology and Application, Tianjin 300222, China

Abstract. In order to solve a series of problems such as complex structure and low training efficiency in traditional deep learning, a fault diagnosis method of rolling bearing based on graph convolution neural network is proposed. Firstly, the convolution layer of neural network is constructed based on graph convolution, and the first-order ChebNet is used to optimize the network model, so as to improve the operation efficiency of the model. Secondly, aggregate the convoluted node information of each layer, and add the features of each layer as the global features of the original graph to achieve effective and accurate feature extraction. Compared with the traditional neural network, the proposed method significantly reduces the complexity and computing time and the network model can still maintain high accuracy when using unbalanced data sets. Through comparative experiments, it is proved that the model has strong feature extraction ability and higher training efficiency, and can still perform well in dealing with the data set with unbalanced sample.

Keywords: Fault diagnosis · Graph convolution · ChebNet · Rolling bearing · Deep leaning · Unbalanced sample

1 Introduction

As the key transmission part of rotating machinery, timely condition monitoring and fault diagnosis of rolling bearing can ensure the safe operation of mechanical equipment [1]. In order to find and judge the fault type in time, how to improve the training efficiency of bearing fault diagnosis process and reliability of diagnosis result is the key of bearing fault diagnosis. As one of the main factors affecting the efficiency and accuracy of bearing fault diagnosis, fault feature extraction[2] has attracted extensive attention.

With the great improvement of computer computing power, machine learning has replaced most artificial feature extraction and become one of the main methods of feature extraction. There are mainly fault diagnosis methods based on support vector machine (SVM) [3] and artificial neural network (ANN) [4]. As an important branch of artificial neural network, deep neural network can effectively extract the deeper features in the

original data and establish the accurate mapping relationship between vibration signal and equipment operation [5]. Common deep learning models include auto encoder (AE) [6], recurrent neural networks (RNN), deep belief network (DBN), and convolutional neural networks (CNN) [7, 8]. Convolutional neural network was first used in image recognition, and then used in equipment fault diagnosis by scholars, and achieved good results. It is widely used in industry for fault diagnosis and condition monitoring of motor, gearbox and bearing equipment, but its accuracy and operation efficiency still have a lot of room for improvement.

Traditional machine learning algorithms usually assume that the data samples are independent and identically distribution, but in real life, things are interrelated and affect each other, there is no completely independent and identically distributed data samples. As a widely existing structure, graph can jointly model the data (Vertex, as the node on the graph) and the connection between data (Edge, as the edge on the graph), so as to achieve more effective and accurate feature extraction. Due to the universality of graph structure, the research of extending deep learning to graphs has attracted more and more attention. The early graph neural networks (GNN) [9] model came into being. Bruna [10] and Defferrard [11] extended the convolutional neural network to the graph and optimized it to produce the graph convolutional neural network. At present, graph convolution neural network is widely used in medical signal processing, molecular structure prediction, computer vision and other fields. For graph convolution network, Kipf et al. [12] proposed the first-order ChebNet. ChebNet can effectively reduce the amount of computation, so it is used for organic molecular structure recognition and so on.

In the actual working condition, the sample data of rolling bearing is usually unbalanced. Because the convolutional neural network is greatly affected by the degree of sample imbalance, it can only be solved by artificially dividing the balanced data set, which makes the samples cannot be fully utilized. However, the graph convolution model based on the first-order ChebNet shows the insensitivity to the unbalanced sample distribution, and can still obtain better training results when using the unbalanced sample data set. Therefore, this paper selects the first-order ChebNet to construct the convolution layer of graph neural network (GCN), and proposes a bearing fault diagnosis method based on GCN. The model has higher accuracy and training efficiency. At the same time, it can effectively deal with the problem of uneven sample distribution in the actual working conditions.

2 Basic Theory

The essence of graph convolution is to generate a new node representation by aggregating the edge information of each node's neighborhood. According to the implementation method, there are mainly methods based on spectral domain model and spatial domain model. At present, graph convolution based on spectral domain has solid theory and technology. Therefore, this type of graph convolution is used for research in this paper.

2.1 Graph Fourier Transform and Graph Convolution

In order to extend the Fourier transform and convolution of convolutional neural network to the graph, the method of replacing the basis function of Laplace operator ($e^{-i\omega t}$) with

the eigenvector of the Laplace matrix of graph is adopted, in which Laplace matrix corresponding to graph is $L = D - A$ (D and A are the degree matrix and adjacency matrix corresponding to the graph respectively). Through eigenvalue decomposition, Laplace matrix can be decomposed into the product of eigenvalues and eigenvectors. The eigenvalue decomposition of Laplace matrix is as follows:

$$L = U \Lambda U^{-1} = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} U^{-1} \quad (1)$$

where $U = (\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n)$ is a matrix composed of unit eigenvectors. Λ is a diagonal matrix composed of eigenvalues of Laplace matrix. Since U is an orthogonal matrix ($UU^T = E$), $L = U \Lambda U^{-1} = U \Lambda U^T$, and U is composed of n linearly independent eigenvectors, that is, $(\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n)$ can form a group of orthogonal bases in the space. When performing the graph Fourier transform, it is equivalent to projecting the input graph signal into the orthogonal space formed by $(\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n)$. To sum up, any vector defined on the graph can be represented by the linear combination of the eigenvectors from its Laplace matrix. Then the above transformation is extended to the matrix form. Combined with the convolution theorem, the convolution of f and graph convolution kernel g can be expressed as:

$$(f * g)_G = U((U^T g) \cdot (U^T F)) \quad (2)$$

Then, $U^T g$ is regarded as a convolution kernel g_θ with learning ability, and the following formula can be obtained:

$$(f * g)_G = U g_\theta U^T F \quad (3)$$

2.2 Rolling Bearing Data Set and Graph Transformation

The data set provided by Case Western Reserve University [13] in the United States was collected by Reliance Electric Motor. Firstly, using EMD on the inner and outer ring rolling elements of motor drive end and fan end bearings generates faults with diameters ranging from 0.007 to 0.028 inches (four kinds, 0.007, 0.014, 0.021 and 0.028 in. respectively). The bearings used by the motor are SKF deep groove ball bearings, with 6205-2RSJEM at the drive end and 6203-2RSJEM at the fan end. Then install each fault bearing on the test bench to operate at a constant speed, to withstand the motor load of 0–3 horsepower (the approximate speed of the motor is 1797–1720 rpm). There are two sampling frequencies, 12 kHz and 48 kHz. In this paper, the data of drive end bearing at 1797 rpm is adopted, the sampling frequency is 12 kHz, and each sample length is 1024 sampling points. The original bearing data is sampled without overlap, and a total of 1305 sample data are obtained. The data set has 10 sample labels, and the corresponding number of samples of each label is shown in Table 1.

The data set is one-dimensional data obtained by resampling the bearing vibration signal rather than graph data type. Therefore, each sample in the data set needs to be

transformed into graph data. Since the complexity of graph convolution operation is $O(N^2)$ (N is the number of nodes), each rolling bearing data sample is segmented to reduce the number of nodes and reduce the amount of operation. The specific method of digitizing dataset graph is given as follows.

Table 1. CWRU rolling bearing data set.

Damage location	Rolling element			Inner ring			Outer ring			Normal
Labels	C1	C2	C3	C4	C5	C6	C7	C8	C9	C0
Damage diameter (inches)	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021	0
Number of samples	118	119	119	118	118	118	119	119	119	238

There are 1024 sampling points in each sample of the data set. Firstly, taking 32 sampling points as a node data and each data sample is divided into 32 nodes. The 32 nodes generate a graph, and the graph data structure is linear. Secondly, the adjacency matrix corresponding to each graph is calculated, and the similarity between the two connected nodes is calculated as the feature of the edge. Finally, for the corresponding sample label on each icon, a total of 1305 graph data can be obtained.

3 Graph Convolution Neural Network

3.1 Symbols and Problem Descriptions

According to the rolling bearing data, the graph convolution neural network performs the graph classification task of a given graph data set $G = \{G_1, G_2 \dots G_n\}$. For any graph G_i , n_i is used to represent the number of nodes, e_i is the number of edges, $A_i \in R^{n_i \times n_i}$ is the adjacency matrix of G_i , and a vector \vec{E}_i with length e_i is used to record the characteristics of its edges, $X_i \in R^{n_i \times f}$ is used to record node information, f is the dimension of the features of X_i , and matrix $Y \in R^{n \times c}$ is used to represent the labels corresponding to each graph (c is the number of sample label types), that is

$$\begin{cases} Y_{ij} = 1; G_i \in class.j \\ Y_{ij} = 0; others \end{cases}$$

According to the above symbols, the problems to be solved can be described as:

- Input: a set of labeled graph training set G_L for learning (generally 80% of data set G is randomly selected as training set G_L).
- Output: prediction results of G_V by trained graph neural network (G_V is the verification set).

3.2 Graph Convolution Neural Network Model Selection

First-order Graph ChebNet (GCN) has been used in many fields and has high recognition accuracy and operation efficiency. Therefore, ChebNet is used to construct convolution layer in this paper. The first-order ChebNet is a network model directly used for graph structure data. According to the first-order ChebNet approximation simplification method of graph convolution, a simple and effective layer propagation method is obtained. Assuming $K = 1$ and $\lambda_{\max} = 2$, the convolution formula can be approximately expressed as

$$x * g_\theta = \theta_0 x - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x \quad (4)$$

In order to reduce network parameters and prevent over fitting, assuming $\theta = \theta_0 = -\theta_1$, the graph convolution can be approximately as follows (θ is the parameter matrix):

$$g_\theta * x = \theta \left(I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right) x \quad (5)$$

Since $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ is an eigenvalue within [0, 2], and the eigenvalue will be used repeatedly during model training, which will lead to numerical divergence and gradient disappearance. Therefore, it is solved by adding a self-ring (introducing a renormalization trick) to the graph, i.e.

$$I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \xrightarrow{\tilde{A}=A+I_N} \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \quad (6)$$

where $\tilde{A} = A + I_N$ is the adjacency matrix with self-connection, and \tilde{D} is the degree matrix ($\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$). Finally, taking it as the input of the activation function, a fast convolution expression can be obtained, i.e.

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (7)$$

where, $\sigma(\cdot)$ is the nonlinear activation function, $H^{(0)} = x$ and $W^k \in R^{d_k \times d_{k+1}}$ is the trainable weight matrix. In order to facilitate parameter adjustment, set the output dimension of all layers $d_{k+1} = d_k = d$. Compared with other graph convolution algorithms, this algorithm has the following advantages.

- Weight sharing and parameter sharing. The parameter matrix of each node is w.
- Local, only the first-order neighbor nodes are aggregated each time.
- The size of receptive field is directly proportional to the number of layers of convolution layer. The more layers, the more sufficient information involved in the operation, the larger the receptive field.
- The complexity is greatly reduced, there is no need to calculate the Laplace matrix and the feature decomposition.

3.3 Overall Neural Network Structure

Figure 1 shows the structure of three-layer graph convolution neural network (GCN3) model. The model mainly has the following three parts.

- Three-layer first-order ChebNet (GCN3). After convolution of each layer of graph, a new node representation will be generated. Through the superposition of three-layer GCN, the neural network of the graph has a third-order receptive field, that is, each point is affected by the third-order adjacent nodes;
- Readout function is used to aggregate node features. Its purpose is to add the convolution features of each layer of graph as the global features of the original graph, that is, gather different levels of graph representations, $Z_i = x_i^1 + x_i^2 + x_i^3$, where x_i^k is the graph representation of the i -th graph after the convolution of the k -layer;
- MLP layer, used for graph classification. The loss function is defined as follow:

$$\hat{Y} = \text{softmax}(\text{MLP}(Z)) \quad (8)$$

$$L = - \sum_{i \in G_L} \sum_{j=1}^c Y_{ij} \log \hat{Y}_{ij} \quad (9)$$

where \hat{Y}_{ij} represents the prediction probability of class j in Graph G_i .

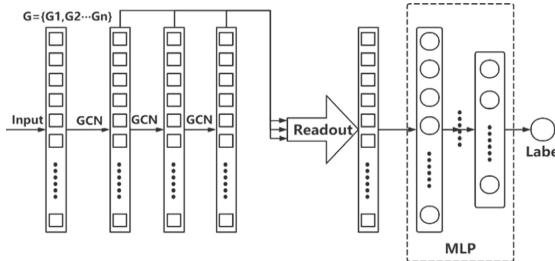
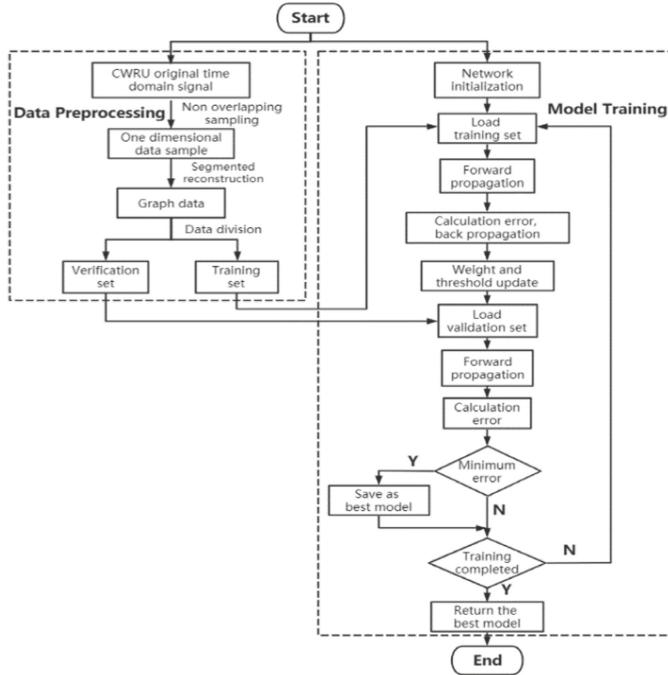


Fig. 1. Three layer graph convolution neural network (GCN3)

The input graph data is subjected to three-layer graph convolution, and then the node representation after each convolution is aggregated through the Readout function to generate a new node representation, which is used for graph classification through the MLP layer. The above model is programmed by Python 3.6 (pytorch 1.3).

4 Experimental Verification

In order to prove that the graph convolution neural network has good performance in training efficiency and recognition accuracy, several groups of comparative experiments are designed in this paper. The data set used in the experiment is the CWRU rolling bearing data set described above. Experiment 1 mainly verified the performance of graph convolution neural network in the case of balanced samples, and Experiment 2 was used to verify the performance in the case of unbalanced samples. The experimental flow chart is shown in Fig. 2.

**Fig. 2.** Experimental flow chart

4.1 Data Sample Division of Experiment 1

The division of balanced samples data is shown in Table 2. To ensure that the distribution of each sample is as uniform as possible and make full use of the data set, except normal, 80 samples are randomly selected from each sample as the training set and 10 as the verification set, with the number of normal samples doubled (The number of normal samples is large, In order to make full use of the sample). The reason for this division of samples data is that in Chapter 4.4, the highest imbalance rate is 30%, $(80 + 10) * 130\% = 117$, less than the minimum sample number 118 in Table 1. In this way the unbalanced data set can be divided without repeated sampling.

Table 2. Division of balanced sample

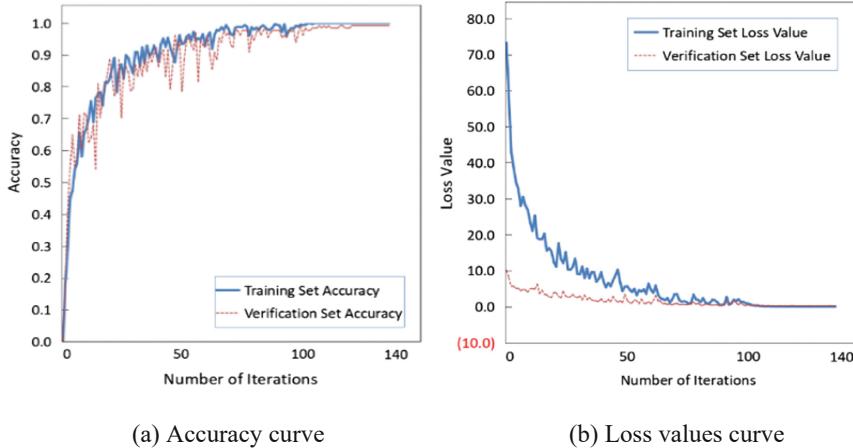
Damage location	Rolling element			Inner ring			Outer ring			Normal
Label	C1	C2	C3	C4	C5	C6	C7	C8	C9	C0
Damage diameter (inches)	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021	0

(continued)

Table 2. (continued)

Damage location	Rolling element			Inner ring			Outer ring			Normal
Samples in training set	80	80	80	80	80	80	80	80	80	160
Samples in validation set	10	10	10	10	10	10	10	10	10	20

4.2 Graph Convolution Neural Network Training of Experiment 1

**Fig. 3.** Accuracy and loss value curve of training set and verification set

The hyperparameters in the experiment are $\text{lr} = 0.003$, batch size = 32, weight decay = $1e-4$, dropout ratio = 0, and the optimizer is LAMB. As can be seen from Fig. 3, when the iteration reaches 110 times, both the accuracy curve and loss value curve begin to stabilize, and the fluctuation is very small in subsequent iterations. The learning rate used in this experiment is $\text{lr} = 0.003$, which is relatively large and leads to the fluctuation amplitude in the early stage of the accuracy curve and the high loss value in the early stage of the loss value curve. Nevertheless, the accuracy and loss value of each experiment tend to be stable (The required number of iterations is slightly different.), which proves that the model has good stability. It can be seen from Fig. 4 that the recognition accuracy of the model has reached 100% after training.

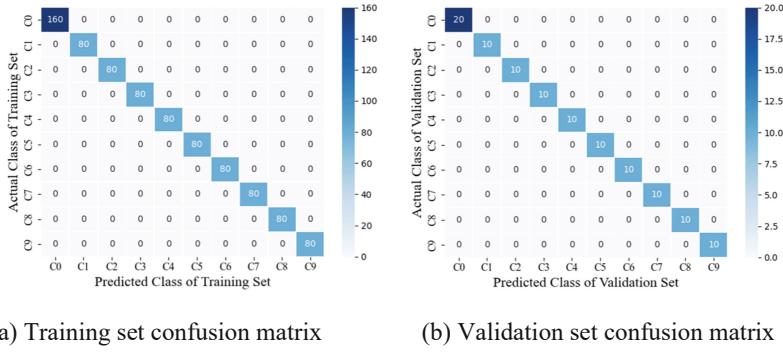


Fig. 4. Confusion matrix of training set and validation set

4.3 Experimental Comparison of Experiment 1

In order to further verify the effectiveness of three-layer graph convolution neural network (GCN3), the model is compared with several typical neural networks (CNN, AlexNet, LeNet5 and ResNet [14]). Because the convolution network cannot use the graph data for training, the two-dimensional rolling bearing data after non overlapping segmentation processing is used for training (the data structure obtained by this method is the closest to the graph data structure described above, and each sample is also 32 * 32, a total of 1024 data. In essence, the graph data structure used in this paper is to segment and recombine the data through the non-overlapped segmentation method), and the same data set as Table 2 is used. Due to space limitations, each convolution network will not be described in detail here. Table 3 is used to show the structure of each network model. Each network model has been tested five times to ensure the accuracy of the experimental result. The comparison results are shown in Fig. 5. Figure 5(a) shows the average accuracy and standard deviation of each network model. Figure 5(b) shows the average time required for 100 training sessions for each network models.

Table 3. Composition of neural network model

	CNN	AlexNet	LeNet5	ResNet
Number of convolution layers	4	5	2	17
Number of pooling layers	2	4	2	2
Number of fully connection layers	2	3	3	1
Activation function	ReLU	ReLU	ReLU	ReLU

It can be seen from Fig. 5(a) that the average recognition accuracy of GCN3 is 99.09%, which is 3.82%, 0.18%, 13.84% and 0.73% higher than CNN, AlexNet, LeNet5 and ResNet models respectively. The standard deviation of recognition accuracy is also

slightly smaller than that of the other four network models, indicating that graph convolution network has stronger feature extraction ability than several typical convolution neural networks.

Because the working principle of graph convolution is different from convolution, the FLOPs of graph convolution is difficult to calculate. Therefore, this paper uses the average time of 100 training under the same operation conditions to compare the operation efficiency of each model. It can be seen from Fig. 5(b) that the time required for 100 training of GCN3 is less than that of other network models, that is, it has higher operation efficiency. The main reason for the high operation efficiency of GCN3 is that GCN3 uses the first-order approximate simplification method and omits the steps of calculating Laplace matrix and eigenvalue decomposition of ordinary graph convolution network, and has a relatively simple network structure.

To sum up, GCN3 reduces the amount of calculation and has high operation efficiency on the premise of ensuring high feature extraction ability.

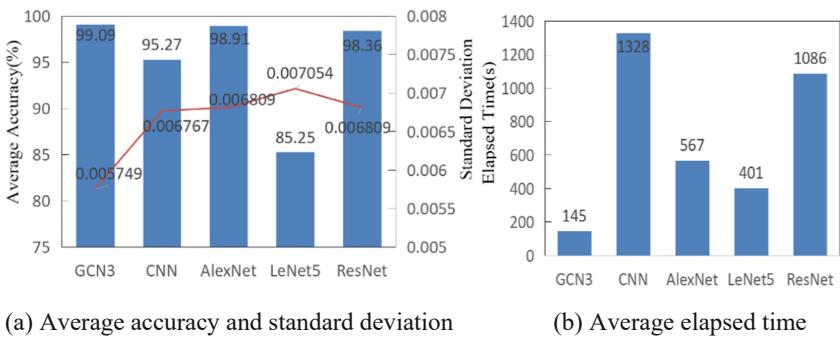


Fig. 5. Results of Experiment 1

4.4 Data Sample Division of Experiment 2

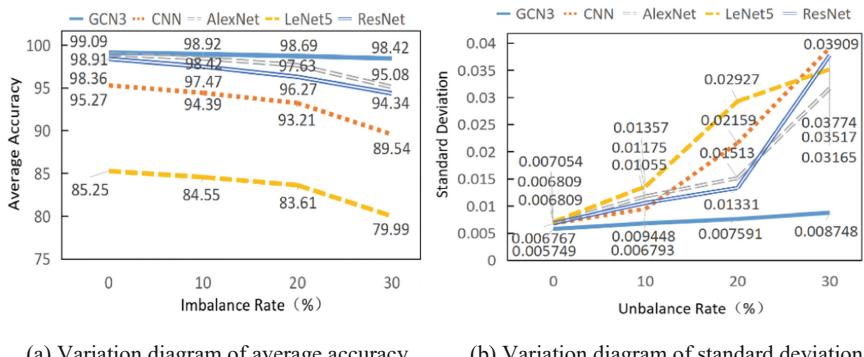
In Experiment 2, the factors affecting the imbalance degree of data samples were added. Based on the number of samples in Experiment 1, the imbalance rates of the three groups of experiments in Experiment 2 increased by 10%, 20% and 30% respectively. The data division of unbalanced samples in the Experiment 2 is shown in Table 4. (The number of samples of each sample is randomly selected from the corresponding range in the table.)

Table 4. Division of unbalanced sample

Damage location		Rolling element			Inner ring			Outer ring			Normal
Label		C1	C2	C3	C4	C5	C6	C7	C8	C9	C0
Damage diameter (inches)		0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021	0
Imbalance rates 10%	Training set	[72,88]									[144,176]
Imbalance rates 20%	Validation set	[9, 11]									[18,22]
	Training set	[64,96]									[128,192]
Imbalance rates 30%	Validation set	[8, 12]									[16,24]
	Training set	[56,104]									[122,208]
	Validation set	[7, 13]									[14,26]

4.5 Experimental Comparison of Experiment 2

In order to more intuitively show that each model is affected by the standard deviation, the average recognition accuracy and standard deviation of each model in Experiment 1 and Experiment 2 are rearranged according to the imbalance rate. The results are shown in Fig. 6.

**Fig. 6.** Variation of accuracy and standard deviation with imbalance rate

It can be seen from Fig. 6(a) that the accuracy of each model decreases in varying degrees with the increase of the imbalance rate. GCN3 is the least affected by the change of the imbalance rate. In the process of increasing the imbalance rate from 0 to 30%, the accuracy decreases by only 0.67%, much less than 3.83% of AlexNet, 4.02% of

ResNet, 5.26% of LeNet5 and 5.73% of CNN. It can be seen from Fig. 6(b) that the standard deviation of each model also increases in varying degrees with the increase of the imbalance rate. The variation range of the standard deviation of gcn3 is much smaller than that of other models. The above two points shows that the feature extraction ability of this model has higher stability than the traditional convolutional neural network and is less affected by the degree of imbalance.

5 Conclusion

In the paper, a fault diagnosis method of rolling bearing based on graph convolution neural network is proposed. The model can diagnosed rolling bearing fault by using bearing vibration signals. After several groups of comparative experiments, the results are as follows.

- The graph convolution network model has high recognition accuracy, the average recognition accuracy can reach 99.09%, which is higher than other network models.
- The graph convolution network model has the advantages of simple structure, small amount of calculation and high training efficiency. Compared with convolutional neural networks, it can obtain better training results faster.
- The graph convolution network model performs well in dealing with the data set with unbalanced sample number. Under the highest imbalance rate, the average recognition accuracy is still 98.42%, and the change of standard deviation is not obvious. It has more stable feature extraction ability than the convolution network model.

In conclusion, the graph convolution neural network model has excellent recognition accuracy and training efficiency, and is not sensitive to the equilibrium degree of the number of samples. It has a wide application prospect in fault diagnosis.

Acknowledgement. This research is a part of the research that is sponsored by the Wuhu Science and Technology Program (No. 2021jc1–6).

References

1. Wang, X., Zheng, J., Pan, H.: Fault diagnosis method for rolling bearings based on MED and autograms. *J. Vibr. Shock* **39**(18), 118–124 (2020)
2. Wang, Z., Yao, L.: Rolling bearing fault diagnosis method based on generalized refined composite multiscale sample entropy combined and manifold learning. *China Mech. Eng.* **31**(20), 2463–2471 (2020)
3. Zhang, C., Chen, J., Guo, X.: A gear fault diagnosis method based on EMD energy entropy and SVM. *J. Vibr. Shock* **29**(10), 216–220 (2010)
4. Benali, J., Fnaiech, N., Saidi, L.: Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. *Appl. Acoust.* **89**, 16–27 (2015)
5. Jiang, H., Shao, H., Li, X.: Deep learning theory with application in intelligent fault diagnosis of aircraft. *J. Mech. Eng.* **55**(7), 27–34 (2019)

6. Sun, W., Shao, S., Zhao, R.: A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement* **89**, 171–178 (2016)
7. Guo, X., Chen, L., Shen, C.: Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. *Measurement* **93**, 490–502 (2016)
8. Wu, Z., Pan, S., Chen, F.: A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4–24 (2020)
9. Scarselli, F., Gori, M., Tsoi, A.C.: The graph neural network model. *IEEE Trans. Neural Netw.* **20**(1), 61–80 (2008)
10. Bruna, J., Zaremba, W., Szlam, A.: Spectral networks and locally connected networks on graphs. Arxiv Preprint. ArXiv <https://arxiv.org/abs/1312.6203> (2013)
11. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. Arxiv Preprint. ArXiv <https://arxiv.org/abs/1606.09375> (2010)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. Arxiv Preprint. ArXiv <https://arxiv.org/abs/1609.02907> (2016)
13. Smith, W.A., Randall, R.B.: Rolling element bearing diagnostics using the case western reserve university data: a benchmark study. *Mech. Syst. Signal Process.* **64–65**, 100–131 (2015)
14. He, K., Zhang, X., Ren, S.: Deep residual learning for image recognition. Arxiv Preprint. ArXiv<https://arxiv.org/abs/1512.03385> (2015)



Research on Bearing Fault Feature Extraction Based on Graph Wavelet

Xin Li¹ and Hui Li^{1,2(✉)}

¹ School of Mechanical Engineering, Tianjin University of Technology and Education,
Tianjin 300222, China
Huili68@163.com

² Tianjin Key Laboratory of Intelligent Robot Technology and Application, Tianjin
300222, China

Abstract. Aim at the problem of large computation and low efficiency of traditional graph convolutional neural networks, a method of extracting bearing fault features based on graph wavelets is proposed. Graph wavelet has the advantages of sparsity and locality, which can provide higher efficiency and better interpretation for graph convolution. Firstly, the fault diagnosis signals of bearing are transformed into ring graph signals. The short bearing fault vibration segment signals are used as nodes, and a set of complete graph signals are formed by edge connection. Secondly, the compressed sparse row sparse matrix of the graph signal is calculated. Finally, the graph wavelet approach is used to extract defect features and classify bearing fault. The experimental results suggest that the graph wavelet-based bearing fault feature extraction approach has good pattern recognition and is a good method for automatic fault feature extraction and pattern identification.

Keywords: Graph wavelet · Graph signal processing · Feature extraction · Fault diagnosis · Bearing

1 Introduction

Deep learning is now widely employed in image processing [1, 2], machine vision, natural language processing [3–5], mechanical defect diagnostics, and other domains as a result of its rapid progress. Convolutional neural networks are commonly employed in mechanical equipment failure diagnostics in deep learning [6]. However, the data processed by the convolutional neural network is Euclidean data, and this type of data processing technology is not enough to completely extract the structural information and feature information of the data [7]. In real life, non-Euclidean structured data is more widely used, for example, social relationship network diagrams, various network topology diagrams, etc. Network graph data consists of nodes (U) and edges (V), which can better represent structurally and feature information. Therefore, scholars have begun to study graph-based convolutional neural networks. Graph convolutional neural networks rely on graph Fourier transform, but graph Fourier transform is computationally intensive and requires feature decomposition. Therefore, it is recommended that the wavelet

basis can be used instead of the Fourier basis, and the wavelet transform can be used to extract the graph signal characteristics. The powerful advantages of graph wavelets are mainly manifested in the aspects of sparseness, high efficiency, and clear features [8].

Signal fault feature extraction [9] and fault categorization [10] are two traditional intelligent fault diagnostic approaches. In recent years, machine intelligence fault diagnostic technology based on deep learning has been successfully used to a variety of equipment defect diagnoses. Li et al. [11], suggested a bearing defect diagnostic approach based on a convolutional neural network and a short-time Fourier transform. A two-dimensional time-spectrogram was created by applying a short-time Fourier transform to the rolling bearing vibration data. Yan et al. [12] used continuous wavelet transform to generate a two-dimensional time-frequency map of the circuit breaker vibration signal, and then down-sampled the time-frequency map and input it to the neural network. However, due to a large amount of computation and many parameters of the convolutional neural network, because the typical machine learning technique expects that the sample data is independent and uniformly distributed, the convolutional neural network application and promotion are harmed. Graph neural networks (GNN) [13] have recently been used in a variety of domains. Spatial domain-based methods and spectral domain-based approaches are the two broad kinds of neural networks. The method based on the spatial domain still follows the convolution method of CNN, in the spatial domain, it explicitly specifies the graph convolution operation. Graph Convolutional Neural Networks (GCN) utilize a normalized Laplacian matrix to aggregate information from neighborhoods. Spectral domain-based methods utilize the graph Fourier transform and the convolution theorem to define convolution. A spectral domain convolution graph neural network (GWNN) [14] is a graph wavelet neural network. Graph wavelet neural networks [15] use graph wavelet basis instead of graph Fourier basis. Since graph wavelet has sparseness, locality, rapidity and low computational cost, it has become an advanced spectral-domain convolution-based graph neural network algorithm.

Taking advantage of graph wavelets, the sparsity, localization, and speed of graph wavelets are used to categorize bearing fault characteristics in this study, which presents a technique for extracting bearing fault features based on graph wavelets. Combining graph wavelets with neural networks can automatically extract fault features and pattern classification. The bearing dataset from Jiangnan University (JNU) [16] is used to verify the efficiency of this approach. For the bearing sampling frequency of 50 kHz, a graph wavelet neural network is utilized, one healthy mode, and three failure modes, respectively, under the condition of three different speeds. A total of 12 types of bearing feature extraction, which can maintain a high recognition rate and is an effective method.

2 Graph Wavelet Based Bearing Fault Feature Extraction

2.1 Graph Signal Processing

The traditional fault feature classification uses the time series of the bearing fault vibration signal to extract the fault feature. However, the graph wavelet bearing fault feature extraction uses the graph signal to perform the feature classification, so the time series needs to be processed by the graph signal.

The time series of bearing fault data is to record the sampling point values of vibration signals in the order of time, so the obtained values are connected as nodes to form a ring diagram. Since the sampling frequency of the bearing dataset of Jiangnan University is 50 kHz and the rotational speed is 600 r/min, 800 r/min, and 1000 r/min, a maximum of 5000 sampling points are required in one cycle. To reduce the computational complexity, a group of 100 sampling points is used as a node. In this paper, 25,000 sampling points of each type of 12 fault types, a total of 300,000 sampling points, composed of 3,000 nodes are connected end to end to form a ring graph.

To calculate the sparse matrix of the graph, record the position row, column, and value of non-zero elements, and store it in CSR format.

Assuming G is an undirected graph, $G = \{V, E, A\}$, where V represents the node set, E represents the edge set, and A represents the adjacency matrix, $A_{i,j} = A_{j,i}$, i and j are the nodes between the nodes connect. Calculating the adjacency matrix A can be expressed as

$$A_{i,j} = \begin{cases} 1 & \text{if } (V_i, V_j) \text{ are connencted on the graph} \\ 0 & \text{if } (V_i, V_j) \text{ aren't connencted on the graph} \end{cases} \quad (1)$$

2.2 Graph Wavelet [14]

Spectral Domain Methods. The graph Laplace matrix is given as follow

$$L = D - A \quad (2)$$

where $D_{i,j} = \sum_j A_{i,j}$. The normalized Laplace matrix is defined

$$L = I_n - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad (3)$$

where I_n is the identity matrix.

The normalized Laplace matrix L has a full set of orthogonal eigenvectors $U = (u_1, u_2, \dots, u_n)$ since it is a real symmetric matrix. It is named as the Laplace eigenvector. The effective value of the non-negative characteristic eigenvalue $\{\lambda_l\}_1^n = 1$ associated with these eigenvectors is the frequency of the graph signal. In the formula, n is the number of nodes. When the eigenvalue is small, it indicates that the signal changes slowly. When the eigenvalue is large, it indicates that the signal changes rapidly.

Graph Fourier Transform. Fourier transform of a continuous signal can be written as

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \quad (4)$$

where $e^{-i\omega t}$ is the basis of the Fourier transform, and the angular frequency is ω .

The eigenvector of the normalized Laplace matrix is used as the foundation for the graph Fourier transform, which defines the signal $x \in R^n$ on the graph G as $\hat{x} = U^T x$, and the graph Fourier transform's inverse is $x = \hat{U}\hat{x}$. According to the convolution theorem, a graph convolution operator can be defined, denoted as $*G$. The convolution kernel is denoted by y , and $*G$ is defined as

$$x * G y = U((U^T y \odot U^T x)) \quad (5)$$

where \odot is the Hadamard product, which means that the corresponding position elements are multiplied. If the vector $U^T y$ is replaced by the diagonal matrix g_θ , matrix multiplication can be used to express the Hadamard product. Using the filter g_θ to filter the signal x , Eq. (5) can be expressed as $U g_\theta U^T x$.

However, implementing graph convolution with Fourier transform has limitations:

- (1) The Fourier basis U is obtained by eigenvalue decomposition using the Laplace matrix, which requires a large amount of calculation.
- (2) When the data dimension is great, the computational efficiency of the graph Fourier transform is low since it must compute the product of the signal x and the dense matrix U .
- (3) The graph convolution after Fourier transform is not limited to the spatial domain, that is, the influence of a node on the signal is not limited to its neighborhood. To effectively solve these limitations, the convolution kernel g_θ can be expanded using the Chebyshev polynomial

$$g_\theta = \sum_{k=0}^{K-1} \theta_k \Lambda^k \quad (6)$$

where K is a hyperparameter that defines the node's neighborhood range distance using the shortest path. $\Lambda = \text{diag}(\{\lambda_i\}_{i=1}^n)$, and $\theta \in R^K$ is a vector of polynomial coefficients. However, the flexibility of the convolution formed on the graph is limited by this polynomial. The lower the value of K , the more difficult the approximation of a diagonal matrix with n free parameters to g_θ . When the value of K is too large, the locality of the graph Fourier transform is also not well resolved. As a result, to address the aforesaid restrictions, this research employs the graph wavelet transform in place of the graph Fourier transform.

Graph Wavelet Transform. The wavelet transform is based on the Fourier transform. However, instead of the Fourier basis, the wavelet basis $\psi(\frac{t-\tau}{a})$ is used. Continuous wavelet transform can be expressed as

$$\text{WT}(a, t) = \frac{1}{\sqrt{a}} f(t) \psi\left(\frac{t-\tau}{a}\right) dt \quad (7)$$

where a is the scale factor, while the translation factor is flat, a controls the expansion and contraction of the wavelet, and the translation factor controls the translation of the wavelet.

The graph wavelet transform is similar to the graph Fourier transform, and the graph signal needs to be projected from the spatial domain to the spectral domain. The graph wavelet transform starts with a collection of wavelets $\psi_s = (\psi_{s1}, \psi_{s2}, \dots, \psi_{sn})$, and ψ_{si} relates to a signal diffused on to network from node i , with s being a scaling parameter. ψ_{si} can be written as

$$\psi_s = U G_s U^T \quad (8)$$

U is the Laplace eigenvector, $G_s = \text{diag}(g(s\lambda_1), \dots, g(s\lambda_n))$ refers to a scaling matrix., and $g(s\lambda_i) = e^{\lambda_is}$.

The definition of the signal x on the graph is $\hat{x} = \psi_s^{-1}x$, and the inverse transformation of the graph wavelet is $x = \psi_s \hat{x}$ based on the graph wavelet. ψ_s^{-1} can be found by substituting $g(-s\lambda_i)$ for $g(s\lambda_i)$ in the thermonuclear equation. Substitute the graph wavelet transform for the graph Fourier transform in Eq. (5), and the graph convolution is obtained as

$$x * Gy = \psi_s((\psi_s^{-1}y) \odot (\psi_s^{-1}x)) \quad (9)$$

In terms of defining graph convolutions, the graph wavelet transform has the following benefits over the graph Fourier transform.

- 1) Efficiency. The graph wavelet has a fast algorithm and does not require Laplace eigenvalue decomposition. To efficiently approximate, Hammond et al. recommended using Chebyshev polynomials, and the complexity is $O(m \times |E|)$, in which $|E|$ denotes the number of sides and m denotes the order of the Chebyshev polynomial.
- 2) High sparsity. For real network matrices, ψ_s and ψ_s^{-1} are sparse, given the sparsity of these networks, the graph wavelet transform is faster than the graph Fourier transform in terms of computing.
- 3) Localized convolution. On the graph, each wavelet corresponds to a signal, it is highly concentrated to the vertex domain and diffuses out from the core node. Therefore, the graph convolution defined in Eq. (7) is limited to the vertex domain.
- 4) Flexible Neighborhoods. Graph wavelets have more flexibility in adjusting node neighborhoods. This method constrains the neighborhood with discrete shortest path distances. This paper uses a continuous method, that is, changing the scaling parameters. Usually, a smaller s value indicates a smaller neighborhood.

Graph Wavelet Neural Network. A graph wavelet neural network is a multilayer convolutional neural network. The m level structure is as follows

$$X_{[:,j]}^{m+1} = h(\psi_s \sum_{i=1}^p F_{i,j}^m \psi_s^{-1} X_{[:,i]}^m) \quad j = 1, \dots, q \quad (10)$$

The wavelet basis is denoted by ψ_s , and ψ_s^{-1} is the graph wavelet transform matrix of the scaling parameter s projecting the signal into the spectral domain, the i column of X^m is $X_{[:,i]}^m$ of dimension $n \times 1$, $F_{i,j}^m$ denotes the diagonal matrix of the spectral domain filter, the non-linear activation function is denoted by h . The host layer converts the input tensor X^m of dimension $n \times p$ into the output tensor X^{m+1} of dimension $n \times q$.

For semi-supervised node classification on graphs, this paper uses a bilayer graph wavelet neural network. The model formulation is as follows:

$$X_{[:,j]}^2 = \text{ReLU}(\psi_s \sum_{i=1}^p F_{i,j}^1 \psi_s^{-1} X_{[:,i]}^1) \quad j = 1, \dots, q \quad (11)$$

$$Z_j = \text{softmax}(\psi_s \sum_{i=1}^q F_{i,j}^2 \psi_s^{-1} X_{[:,i]}^2) \quad j = 1, \dots, c \quad (12)$$

The cross-entropy of all tokens is the loss function, that is

$$\text{Loss} = - \sum_{l \in L} \sum_{i=1}^c Y_{li} \ln Z_{li} \quad (13)$$

The number of classes in the node categorization is given by c and the prediction of dimension $n \times c$ is Z . The set of labelled nodes is y_L with $Y_{li} = 1$ if the subscript of node l is i and $Y_{li} = 0$ otherwise. Gradient descent is used to train the weights F .

Each layer's computational complexity is given by $(n \times p \times q)$ in Eq. (10), in which n is the number of nodes, p is the number of features per vertex in the current layer, and q is the number of features per vertex in the next layer. Traditional convolutional neural networks learn convolutional kernels for each set of input features, as a result, a huge number of parameters are generated, and parameter learning typically necessitates a big amount of training data. To overcome this issue, this research divides each layer of the graph wavelet neural network into two parts: the feature transform and the graph convolution, which are expressed as feature transform and graph convolution, respectively.

$$\text{feature transform : } X^m' = X^m W \quad (14)$$

$$\text{graph convolution: } X^{m+1} = h(\psi_s F^m \psi_s^{-1} X^m) \quad (15)$$

In which $W \in R^{p \times q}$ is the eigentransform parameter matrix, the eigen matrix following the eigen transform is $X^{m'}$ of dimension $n \times q$, F^m is the nonlinear activation function, and h is the kernel function of the diagonal matrix of the graph convolution.

The computational complexity changes from $O(n \times p \times q)$ to $O(n + p \times q)$ when the feature transform is decoupled from the graph convolution.

Graph Wavelet Based Bearing Fault Feature Extraction. The graph wavelet based bearing fault diagnosis feature extraction steps include data pre-processing, feature extraction of bearing faults, classification, and visualization. Data pre-processing is the

conversion of sample data into graph data. As the sample data is a time series, it is necessary to form a set of rings of graph data with a short segment of continuous data as nodes and to calculate the Compressed Sparse Row (CSR) sparse matrix and adjacency matrix as input data. Feature extraction is mostly used to extract bearing fault characteristics and classify fault kinds automatically. The classification visualization is a visualization of the output results and allows the classification effect to be checked visually.

The flow chart for graph wavelet bearing fault feature extraction is shown in Fig. 1.

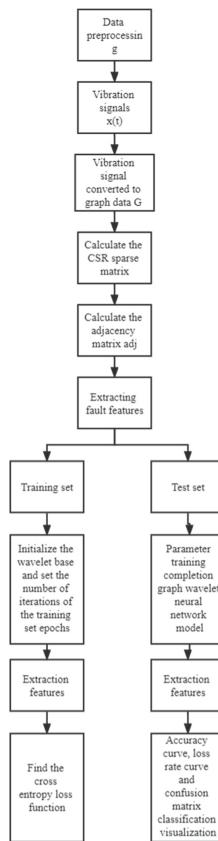


Fig. 1. Flow chart of graph wavelet bearing fault feature extraction

The main steps in graph wavelet bearing fault feature extraction are as follows.

- 1) Convert the vibration signal x into graph data G .
- 2) Calculate the sparse matrix CSR.
- 3) Calculate the adjacency matrix adj.
- 4) Divide the training set, test set, and validation set.
- 5) Input the divided data into the input layer of the graph wavelet neural network, train the graph wavelet neural network parameters and perform feature extraction.

- 6) Visualisation of the bearing fault classification.

3 Graph Wavelet Based Bearing Fault Feature Extraction

3.1 Experimental Data

The rolling bearing dataset from Jiangnan University (JNU) was utilized for experimental testing of the graph wavelet based bearing defect feature extraction technique. The JNU bearing dataset motor vibration signal, the sampling frequency is 50 kHz, and the fault bearing is a deep groove ball bearing. In this paper, the fault is divided into 12 categories according to the location of the damage which is divided into damage (tb), inner ring damage (ib), outer ring damage (ob), and normal bearings (n), according to the speed which is divided into 600r/min, 800r/min and 1000r/min. According to various bearing damage which are marked as ib600, ib800, ib1000, ob600. The bearings are marked as ib600, ib800, ob1000, tb600, tb800, tb1000, n600, n800, n1000, as shown in Table 1. Each type of data needs to collect 25,000 sampling points, and 100 sampling points as a group as a node of the graph data, each type of data corresponds to 250 nodes, a total of 3,000 nodes ring connected to form the graph data, verification set, test set according to 7:2:1 division, as shown in Table 1. In the training process, the labels for each class of faults need to be in the form of one-hot encoding.

Table 1. Experimental data set

Location of injury		Rolling body			Inner circle			Outer circle			n		
Labels		1	2	3	4	5	6	7	8	9	10	11	12
	tb600	tb800	tb1000	ib600	ib800	ib1000	ob600	ob800	ob1000	n600	n800	n1000	
RPM (r/min)	600	800	1000	600	800	1000	600	800	1000	600	800	1000	
Dataset	Training	175	175	175	175	175	175	175	175	175	175	175	175
	Validate	50	50	50	50	50	50	50	50	50	50	50	50
	Testing	25	25	25	25	25	25	25	25	25	25	25	25

3.2 Analysis of Experimental Results

The whole network consists of 2 convolutional layers, using 128 hidden units as the output dimension of the first convolutional layer, uses an Adam optimizer, a maximum Chebyshev polynomial count of 3, and an initial learning rate of 0.01.

The activation function is ReLU, and to improve computational efficiency. Set to 0 when ψ_s and ψ_s^{-1} are less than the threshold. The wavelet coefficient s = 1.0, and the sparsity threshold t = 0.0001.

The dataset maximum number of iterations (epochs) is set at 600, with the validation loss remaining constant for 30 consecutive epochs, otherwise, the training will be terminated. Experimental platform configuration: Windows 10 64-bit operating system,

Intel(R) CoreTMi7-8750H CPU @2.20GHz, program running environment is Tensorflow1.2. Figure 2 shows the training set and test set accuracy (accuracy) and target error (loss) curves after the training was finished.

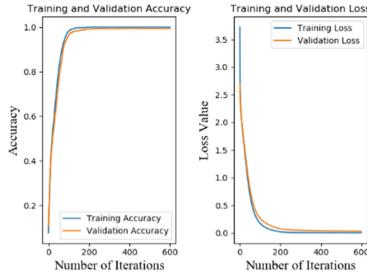
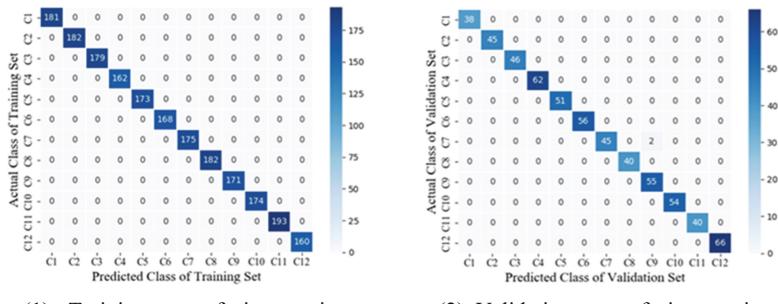


Fig. 2. Accuracy and target loss curves

As can be seen from Fig. 2, the loss value converges to 0 at 250 iterations. The accuracy of the training and test sets converges to 100% at around 200 iterations, and the model stabilizes after 400 iterations, with the final accuracy of the training and test sets of the model reaching 100% and 99%.



(1) Training set confusion matrix

(2) Validation set confusion matrix

Fig. 3. Confusion matrix of the training and validation sets

The confusion matrix was used to illustrate the classification results of the training and test sets after 600 iterations, as shown in Fig. 3. Fig. 3(1) is the confusion matrix of the training set, the validation set's confusion matrix is shown in Fig. 3(2), where the horizontal and vertical axes represent the anticipated and actual fault kinds, respectively, observing the diagonal of the confusion matrix, the results show that the bearing fault feature extraction method of graph wavelet has separated the different fault types of The samples were well separated, the neural network made slight errors in judging C7 faults, two were classified into C9 category, and overall a good classification result was obtained.

4 Performance Comparison of GWNN and GCN

4.1 Experimental Results Analysis

Graph wavelet neural networks (GWNN) and graph convolutional neural networks (GCNN) are empirically compared to evaluate the usefulness of graph wavelet neural networks for node classification (GCN).

The parameters of a graphical convolutional neural network (GCN) are as follows. With an initial learning rate of 0.0001 and a maximum number of Chebyshev polynomials of 3. The network is trained as a graphical convolutional neural network with 128 hidden units using the Adam optimizer. The activation function is ReLU, and the maximum number of iterations (epoch) is 60,000.

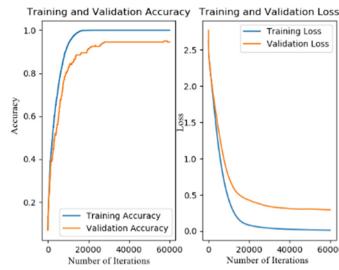
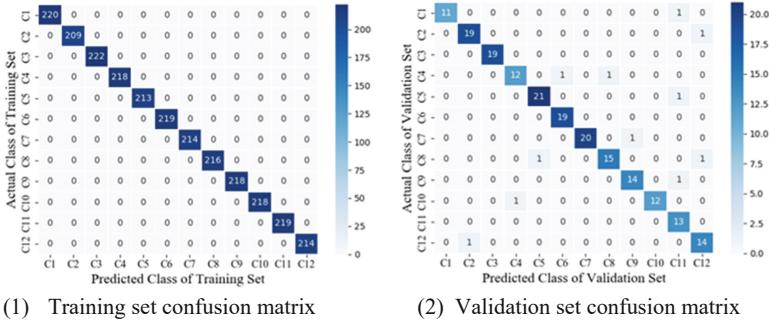


Fig. 4. Accuracy and target loss curves



(1) Training set confusion matrix

(2) Validation set confusion matrix

Fig. 5. Confusion matrix of the training and validation sets

The weight of L_2 loss on the embedding matrix is 0.00001. Meanwhile, the validation loss was kept from decreasing for 30 consecutive epochs, otherwise, the training would be terminated. The training set and test set accuracy (accuracy) and target error (loss) curves are displayed in Fig. 4 after the training is finished.

The training set and test set accuracy (accuracy) and target error (loss) curves are displayed in Fig. 4 after the training is finished, and the model stabilizes after 50,000 iterations, with the final accuracy of the training and test sets of the model reaching 100% and 92%.

The confusion matrix is used to depict the classification results after 60,000 iterations of the training and test sets, as shown in Fig. 5. The results reveal that when using the graph convolution technique to extract bearing fault features, samples of different fault kinds were separated, but that one data of each type was classified into the other types in the validation set.

4.2 Comparative Analysis of Experimental Results

Table 2 compares and displays the outcomes of the GWNN and GCN experiments.

Table 2. Comparison of experimental results between GWNN and GCN

Network type	Dataset	Training set acc	Validation set acc	Number of iterations
GWNN	JNU	100%	99.2%	600
GCN	JNU	100%	92.4%	60000

Using the wavelet transform instead of the Fourier transform, it can be seen from Table 2 that GWNN is ahead of GCN, and the correct rate of the validation set is 6.8% higher for GWNN compared to GCN. There are two possible explanations for this improvement:

- 1) Because the vertex domain convolution is non-local, the GCN's feature spread is not confined to adjacent nodes.
- 2) GWNN's scaling parameters are more adaptable, enabling the diffusion range to be modified and making it more suited to a variety of datasets. Since GWNN has enough degrees of freedom to learn convolutional kernels, while GCN uses a limited number of degrees of freedom, GWNN has higher accuracy. The number of iterations between GWNN and GCN shows that GWNN has 600 iterations, while GCN requires up to 60,000 iterations, so GWNN is efficient compared to GCN.

In addition to the high accuracy and efficiency of the GWNN, the wavelet transforms with a locally sparse transform matrix is also sparse in the spectral and spatial domains. Because the dense matrix U must be multiplied by the graph signal x in the graph Fourier transform, whereas the graph wavelet transform is multiplied by the sparse matrix ψ_s , as a result, the graph wavelet transform has fewer non-zero elements in the transform matrix and the projected signal than the graph Fourier transform. The graph wavelet's sparsity speeds up computation and better reflects the topology of the neighborhood centered on each node.

5 Conclusion

The network graph corresponds to the realistic structure of everything connected, which is closer to real life, so it is more effective in retaining minute features, and the graph wavelet neural network will have better feature extraction capability when dealing with fault feature extraction of bearings. The graph wavelet neural network is proposed instead of the graph Fourier transform, and it has three advantages: (1) The graph wavelet is locally sparse; (2) A time-saving approach is the graph wavelet transform; (3) The convolution takes place just in the vertex domain. These benefits make the entire training process more efficient and straightforward. In the extraction of bearing defect features, the graph wavelet neural network produced good feature classification results.

Acknowledgement. This research is a part of the research that is sponsored by the Wuhu Science and Technology Program (No. 2021jc1–6).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* **25**(2), 1097–1105 (2012)
2. Farabet, C., et al.: Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1915–1929 (2013)
3. Hirschberg, J., Manning, C.D.: Advances in natural language processing. *Science* **349**(6245), 261–266 (2015)
4. Sun, S., Luo, C., Chen, J.: A review of natural language processing techniques for opinion mining systems. *Inf. Fusion* **36**, 10–25 (2017)
5. Young, T., Hazarika, D., Poria, S., et al.: Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13**(3), 55–75 (2018)
6. Dai, X.: Application of artificial intelligence in industrial robotic systems. *China-Arab Sci. Technol. Forum* **1**, 99–101 (2021)
7. Zheng, W.: Research on Representation Learning Algorithm Based on Graph Wavelet Neural Network, pp. 1–65. Anhui University, Hefei (2021)
8. Zhao, Z., Wu, S., Qiao, B., et al.: Enhanced sparse period-group lasso for bearing fault diagnosis. *IEEE Trans. Ind. Electron.* **66**(3), 2143–2153 (2019)
9. Wang, S., et al.: Matching synchrosqueezing wavelet transform and application to aeroengine vibration monitoring. *IEEE Trans. Instrum. Meas.* **66**(2), 360–372 (2017)
10. Sun, C., Ma, M., Zhao, Z., et al.: Sparse deep stacking network for fault diagnosis of motor. *IEEE Trans. Ind. Inf.* **14**(7), 3261–3270 (2018)
11. Li, H., Zhang, Q., Qin, X., et al.: Bearing fault diagnosis method based on short-time Fourier transform and convolutional neural network. *Vibr. Shock* **37**(19), 124–131 (2018)
12. Yan, R., Lin, W., Gao, S., et al.: Analysis of circuit breaker fault diagnosis based on wavelet time-frequency diagram and convolutional neural network. *Vibr. Shock* **39**(10), 198–205 (2020)
13. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations, pp. 1–14 (2016)
14. Xu, B., et al.: Graph wavelet neural network. In: Proceedings of the International Conference on Learning Representations, pp. 1–13 (2019)

15. David, K.: Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* **30**(2), 129–150 (2011)
16. Li, K.: School of Mechanical Engineering. Jiangnan University, Wuxi (2019). <http://mad-net.org:8765/explore.html?t=0.5831516555847212>



Correntrogram: A Robust Method for Optimal Frequency Band Selection to Bearing Fault Detection

Hui Li^(✉), Ruijuan Wang, and Yonghui Xie

School of Mechanical and Electrical Engineering,
Weifang Vocational College, Weifang 262737, China

Huili68@163.com

Abstract. Correntropy includes not only the second-order statistics of the signal, but also the higher-order statistics of the signal. Therefore, correntropy is an effective tool to deal with nonlinear and non-Gaussian signals. In order to solve the problem that it is difficult to select the optimal frequency band of bearing fault vibration signal under the interference of Gaussian and non-Gaussian Noise, a new optimal frequency band selection method is proposed, which is named as Correntrogram. Firstly, the correntropy of the signal is calculated. Then correntropy is decomposed into multiple frequency bands using the 1/3-binary tree structure and the optimal frequency band is selected according to the L_2/L_1 norm. Finally, the squared envelope spectrum of the optimal frequency band is calculated and bearing fault characteristics frequency can be accurately identified. The results of simulation and experiment show that Correntrogram can correctly select the optimal frequency band of bearing fault vibration signal under the interference of Gaussian and non-Gaussian noise, which has good robustness, and its performance is better than that of traditional Kurtogram.

Keywords: Correntropy · Bearing · Fault detection · Kurtogram · Optimal frequency band

1 Introduction

Gearbox is one of the key parts of rotating machinery equipment. Because of its complex structure and bad working environment, it causes frequent failures. The transmission path of vibration signals collected from gearbox is complex and changeable, which makes it difficult to extract fault features effectively by traditional methods. In order to analyze the time-varying non-stationary signals more effectively, the time-frequency domain analysis method comes into being. By combining the time-domain analysis with the frequency-domain analysis, not only can the frequency-domain characteristics of the signals be described, the time-dependent characteristics of the frequency domain can also be characterized. Common time-frequency analysis methods include empirical mode decomposition (EMD), empirical wavelet transform (EWT), Wigner-ville distribution (WVD) and Short-time Fourier transform (STFT) [1, 2]. EMD method can

adaptively decompose the signal according to its time scale characteristic, and it does not need to set up the basis function in advance, so it is very suitable to deal with the non-linear and non-stationary signal such as the vibration signal of gearbox, it has a high signal-to-noise ratio. Although EMD has achieved good results in gearbox fault diagnosis, it is still not immune to complex vibration signals, and the problem of end-point effect and mode aliasing often occurs [3]. WVD is a kind of quadratic transformation, which can be regarded as the Fourier transform of the instantaneous correlation function of the signal, and can describe the time-varying characteristics of the signal well, but it is easily disturbed by the cross terms, the application of this method is limited. Kurtosis can describe the impact characteristics of fault pulse components in vibration signals, and is one of the commonly used time domain indicators for fault diagnosis of mechanical rotating equipment. Using kurtosis in the frequency domain, the resonance frequency band of the signal can be obtained and the signal can be bandpass filtered to achieve the effect of noise reduction. To solve this problem, Antoni [4–6] put forward the concept of spectral kurtosis (SK), which determines the optimal frequency band of the fault by calculating the kurtosis of amplitude of the signal, the center frequency and bandwidth are obtained to filter the original signal. Xu et al. [7] used variational mode decomposition (VMD) to decompose the vibration signal of the gearbox into a series of modal components, and then filtered the modal components with the largest correlation coefficient with the original signal for fast spectral kurtosis processing, so as to achieve filtering and noise reduction. Finally, the filtered signal is analyzed by envelope demodulation to complete the fault diagnosis. However, the collected vibration signals are often affected by random impact or interference noise, so the diagnosis using spectral kurtosis may lead to an incorrect resonance band and poor filtering effect. Antoni [8] merges the negative entropy of time domain spectrum and the negative entropy of frequency domain spectrum, and puts forward the information graph method, which considers the impulse and cyclostationarity of vibration signal synthetically, it is widely used in the field of mechanical fault diagnosis. When a part of the gearbox is damaged locally, the impulsive signal produced by the gearbox will be distorted and mixed in different degrees during the path transmission, the vibration signals collected by sensors often contain noise components, which leads to low signal-to-noise ratio, weak fault characteristics and difficult to extract.

Although Kurtogram is widely used in the field of fault diagnosis, Kurtogram also has some shortcomings. First, the number of repeated shock pulses seriously affect the kurtosis value. The kurtosis value of a single shock pulse is the largest, and the kurtosis value decreases with the number of repeated shock pulses, the frequency band selected according to the principle of maximum kurtosis is not necessarily the optimal one, so it is easy to choose the pseudo-optimal one in fault feature extraction. Secondly, Kurtogram is seriously affected by noise, and kurtosis decreases rapidly with the decrease of signal-to-noise ratio (SNR). Therefore, when SNR is very low, especially when the signal contains non-gaussian noise, it is difficult to obtain the ideal diagnosis effect.

Correntropy takes into account not only the statistical properties of random time series, but also the time structure of random time series [9, 10]. Correntropy is a kernel-based measure of local similarity among random variables. It is a generalization of

traditional correlation function and contains the characteristics of traditional correlation function, but its performance is better than that of traditional correlation function. Correntropy can not only suppress Gaussian noise, but also non-Gaussian noise. Therefore, correntropy based on kernel function provides a new robust processing method for non-Gaussian noise, it has been widely concerned in the field of communication signal processing. In this paper, a novel Kurtogram named Correntrogram is proposed to improve the conventional Kurtogram. The main aim of Correntrogram is to suppress the Gaussian and non-Gaussian noise and improve the robustness of the optimal frequency band selection in order to effectively identify the fault characteristics of bearings. Firstly, the correntropy of the signal is calculated, then the correntropy is decomposed into multiple frequency bands and the optimal frequency band is selected according to the L_2/L_1 norm. Finally, the squared envelope spectrum of the optimal frequency band is calculated. The bearing fault characteristics frequency is identified according to squared envelope spectrum.

The rest of the arrangements are outlined as follows. The basic theory of correntropy, definition of spectral L_2/L_1 norm and main steps of Correntrogram for bearing fault detection are proposed in Sect. 2. Section 3 verifies the effectiveness of proposed method using a simulative signal. Section 4 validates the advantages of Correntrogram using bearing inner race fault signal. Conclusions are drawn in Sect. 5.

2 Correntrogram for Bearing Fault Detection

2.1 The Basic Theory of Correntropy

For random processes $x(t)$, auto-correntropy can be defined as

$$V_x^\sigma(\tau) = E\{\kappa_\sigma[x(t), x(t + \tau)]\} \quad (1)$$

where τ is time lag, $E(\cdot)$ is expected mean operator, $\kappa_\sigma(\cdot)$ is Mercer kernel function. The Gaussian kernel function is selected, which is expressed as

$$\kappa_\sigma[x(t), x(t + \tau)] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[x(t) - x(t + \tau)]^2}{2\sigma^2}\right\} \quad (2)$$

where σ is kernel size, $\exp(\cdot)$ is exponential function.

Correntropy takes into account not only the statistical properties of random time series, but also the time structure of random time series. Correntropy is a kernel-based measure of local similarity among random variables. It is a generalization of traditional correlation function and contains the characteristics of traditional correlation function, but its performance is better than that of traditional correlation function [9]. The advantages of correntropy are given as follow.

- 1) Correntropy includes not only the second-order moment of random variable amplitude, but also the higher-order moment of random variable amplitude, so correntropy can describe the nonlinear characteristic of signal more effectively.

- 2) Correntropy makes use of kernel trick, which can efficiently perform inner product operation and save computing time.

Correntropy can suppress not only Gaussian noise but also non-Gaussian noise effectively. Therefore, the correntropy based on kernel function provides a new robust solution for non-Gaussian noise.

2.2 Definition of Spectral L₂/L₁ Norm and Signal Decomposition

Spectral L₂/L₁ norm is widely used in sparse signal processing. Compared with spectral kurtosis, it can emphasize the transient impulse and sparsity of the signals at the same time. It also can more efficiently reduce the interference of noise. Therefore, the spectral L₂/L₁ norm is applied as an indicator for optimal frequency band selection. The larger spectral L₂/L₁ norm is, the more fault characteristic information indicates. Spectral L₂/L₁ norm is defined as

$$K_k^i = \frac{\|c_k^i(n)\|_2}{\|c_k^i(n)\|_1} = \sqrt{\frac{\sum_{n=1}^N |c_k^i(n)|^2}{\sum_{n=1}^N |c_k^i(n)|}} \quad (3)$$

where $c_k^i(n)$ is the filter coefficients of the i th filter, $i = 0, 1, \dots, 2^k - 1$, at the k th level of the filter bank of the 1/3 binary tree. $k = 0, 1, 2, \dots, N_{level}$, N_{level} is the maximum number of layers of signal decomposition. Please refer to Ref.[6] for details.

According to Ref.[6], the center frequency of each frequency band is expressed as

$$f_i = (i + 0.5)2^{-k-1} \quad (4)$$

The bandwidth of each frequency band is given as

$$BW_k = 2^{-k-1} \quad (5)$$

2.3 Main Steps of Correntrogram for Bearing Fault Detection

The main process of Correntrogram for bearing fault detection is as follows:

- 1) To calculate the correntropy according to Eq. (1).
- 2) To decompose correntropy into multiple frequency bands using the 1/3-binary tree structure and the optimal frequency band is selected according to Eq. (3).
- 3) To calculate the squared envelope spectrum of filtered signal and bearing fault characteristics frequency can be identified.

3 Simulative Signal Analysis

In order to validate the performance of Correntrogram, a simulative signal $x(t)$ is designed to simulate the bearing fault signals generated in the working process of rotating machine. Simulative signal $x(t)$ is modeled according to Ref.[11]:

$$x_1(t) = \sum_{i=1}^N A_i \times \Theta(t - t_i) e^{-C_i(t-t_i)} \times \sin[2\pi f_{bi}(t - t_i) + \theta_{bi}] \quad (6)$$

$$x(t) = x_1(t) + n_1(t) + n_2(t) \quad (7)$$

where A_i is the transient pulse amplitude, C_i is the damping attenuation factor, t_i is the impact duration, θ_{bi} is the initial phase and f_{bi} is the resonance frequency of the mechanical system; $n_1(t)$ is zero mean Gaussian noise and $n_2(t)$ is impulsive noise. The function $\Theta(t - t_i)$ is used to specify when the shock occurs.

The fault characteristic frequency of the bearing outer ring is $f_{outer} = 110$ Hz. The natural vibration frequency of the mechanical system is $f_b = 3000$ Hz, the transient impact amplitude = 0.5, the sampling frequency $f_s = 10000$ Hz and the number of signal sampling points is $n = 4000$.

The impulsive noise $n_2(t)$ is modeled as single degree freedom vibration system. The resonant frequency f_l of impulsive noise is equal to 3200 Hz and attenuation frequency is 40 Hz.

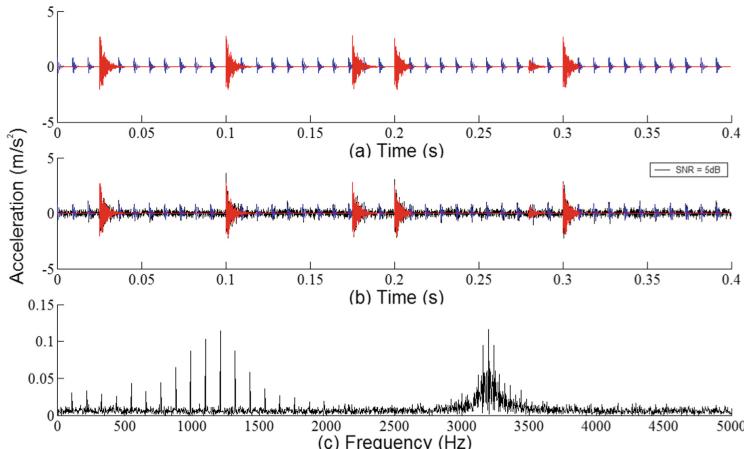


Fig. 1. The simulative signal with two resonant frequencies bands. (a) Bearing fault signal (blue) and impulsive noise (red); (b) Bearing transient fault signal, impulsive noise and Gaussian noise; (c) Their Fourier spectrum. (Color figure online)

The simulative signal is displayed in Fig. 1. Figure 1(a) displays the six impulsive noise located at sampling point 250, 1000, and so on, which have larger magnitude than that of the simulative bearing defect transients signal. Figure 1(b) shows the simulative

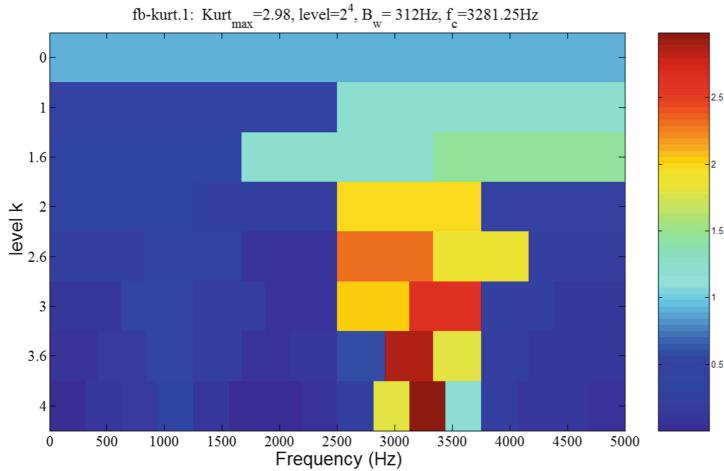


Fig. 2. Kurtogram of the simulative signal

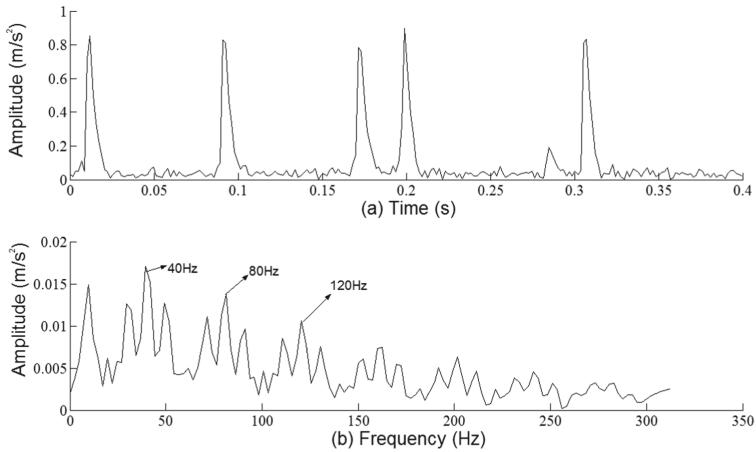


Fig. 3. (a) Amplitude envelope of the filtered signal with maximized Kurtogram; (b) Squared envelope spectrum

signal after adding Gaussian noise, which the SNR is 5dB. Figure 1(c) exhibits their Fourier spectrum.

The Kurtogram is firstly used to select the optimal frequency band of the simulative displayed in Fig. 1(b) and the results are displayed in Fig. 2. It can be noted that at central frequency 3281.25 Hz has the largest kurtosis value along all of the filter frequency bands. The corresponding envelope and the squared envelope spectrum are shown in Fig. 3. Figure 3(b) exhibits that the filtered signal is exactly the impulsive noise component, which the frequency is exactly 40 Hz. Figure 3 demonstrates that the traditional Kurtogram can not correctly select the optimal frequency band for bearing failure due to impulsive noise.

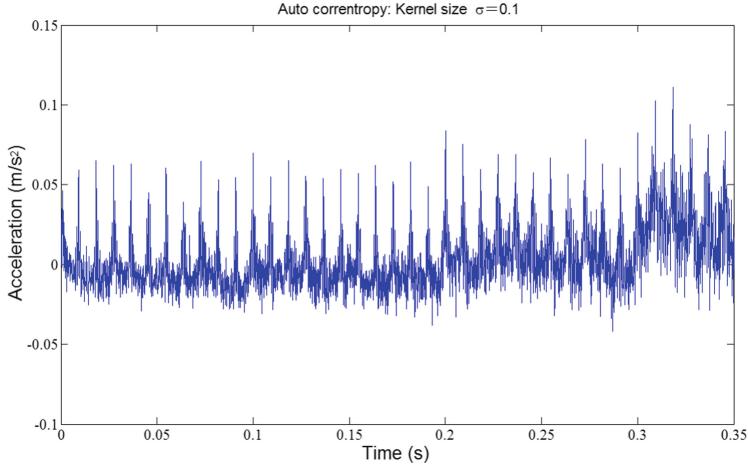


Fig. 4. Correntropy of the simulative signal

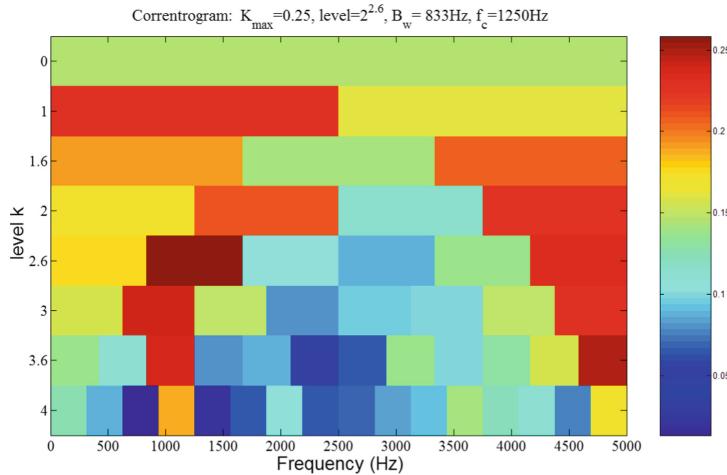


Fig. 5. Correntrogram of the simulative signal

For comparison, the proposed method is applied to simulative signal displayed in Fig. 1(b). The correntropy is calculated according to Eq. (1) and the corresponding results are shown in Fig. 4. From Fig. 4, the periodic transient impulsive shocks are observed. It is shown that the correntropy can extract the weak bearing fault information from the signal under the interference of Gaussian and non-Gaussian noise.

The Correntrogram is used to select the optimal frequency band of the correntropy displayed in Fig. 4 and the results are displayed in Fig. 5. It can be noted that at central frequency 1250 Hz has the largest spectral L_2/L_1 norm along all of the filter frequency bands. The corresponding envelope and the squared envelope spectrum are shown in Fig. 6. Figure 6(b) exhibits that the filtered signal is exactly the bearing fault component,

which the frequencies is corresponding to bearing fault characteristic frequency (f_{outer}) and its higher harmonics. This means that Correntrogram can determine the correct position of the carrier frequency of the transient impulsive shocks produced by defected bearing. By comparison, Correntrogram is superior to traditional Kurtogram methods in analyzing the simulated bearing fault signal.

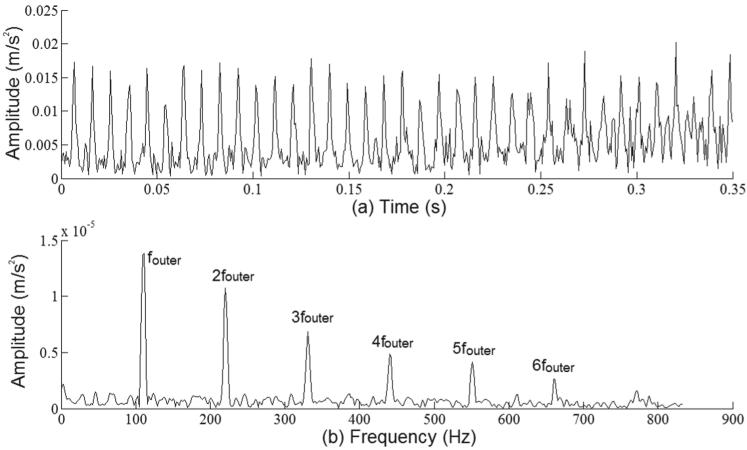


Fig. 6. (a) Envelope of the filtered signal with maximized Correntrogram; (b) Squared envelope spectrum

4 Experimental Verification of Bearing Fault Detection

The rolling bearing inner race fault signal is selected for investigation in this part to test the accuracy of the Correntrogram in real applications. Meanwhile, the traditional Kurtogram method is applied to process the same experimental signal for comparison. In the experiment, the defected bearing is a deep groove ball bearing of type 6205 with electrical discharge machining damage to the inner race. The depth of this defect is up to 0.034 in. and the width of the defect is 0.007 in. The sampling frequency is 12000 Hz and the data sampling points is 4000. The bearing inner race rotating frequency (f_r) is 29.17 Hz. Then, the bearing fault characteristic frequency is computed as $f_{\text{inner}} = 157.96$ Hz.

Figure 7(a) depicts the time-domain waveform of the bearing inner fault vibration signal. Although the time domain waveform can partly express the periodic transient impulsive shocks of vibration signal, it cannot identify the period to reflect the bearing inner fault information. Figure 7(b) depicts the Fourier spectrum of the bearing inner fault vibration signal. The bearing inner fault characteristic frequency is buried by other frequencies and noise in Fig. 7(b). As a result, the bearing inner race fault information cannot be directly obtained from Fourier spectrum.

In order to verify the effectiveness of Correntrogram, the proposed method is applied to bearing inner defect vibration signal shown in Fig. 7(a). The correntropy is calculated

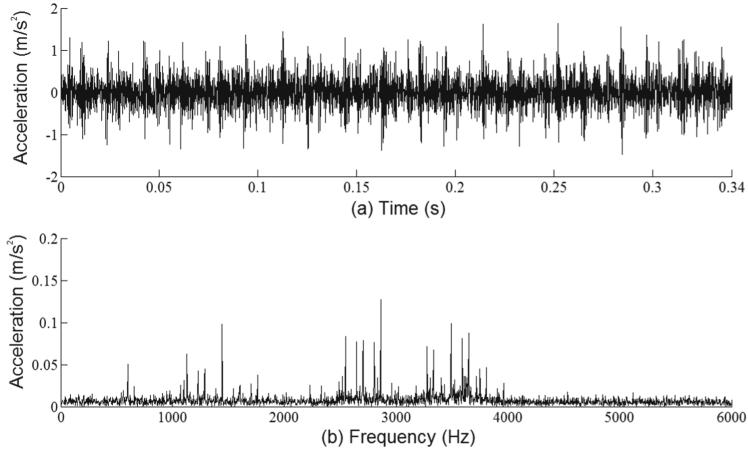


Fig. 7. (a) Bearing inner fault signal; (b) Fourier spectrum

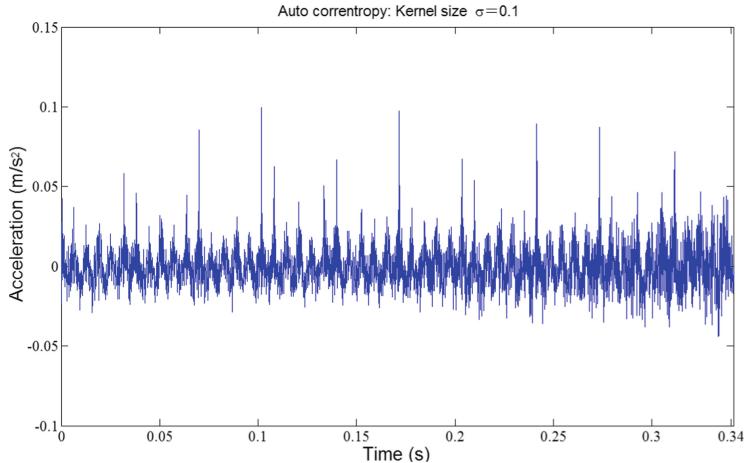


Fig. 8. Correntropy of bearing inner fault signal

according to Eq. (1) and the corresponding results are shown in Fig. 8. From Fig. 8, the periodic transient impulsive shocks are observed. The interval between the two adjacent transient pulses is exactly equal to the characteristic period of bearing inner ring fault. It is demonstrated that the correntropy can extract the weak bearing fault information from the signal perturbed by other frequencies and noise.

Then, the bearing inner race fault vibration signal is dealt with using the Correntrogram and the resulting results are exhibited in Fig. 9. As seen in Fig. 9, Level 1.6 has the frequency band with the highest spectral L_2/L_1 norm, corresponding to its center frequency and bandwidth being 3000 Hz and 2000 Hz respectively. The components extracted by the optimal frequency band and its squared envelope spectrum are shown

in Fig. 10(a) and Fig. 10(b) respectively. The bearing inner race fault characteristic frequency (f_{inner}) and its higher harmonics can be easily observed from the squared envelope spectrum. This demonstrates that the Correntrogram can reliably identify bearing fault characteristic information, allowing it to be used for bearing inner race fault detection and pattern recognition.

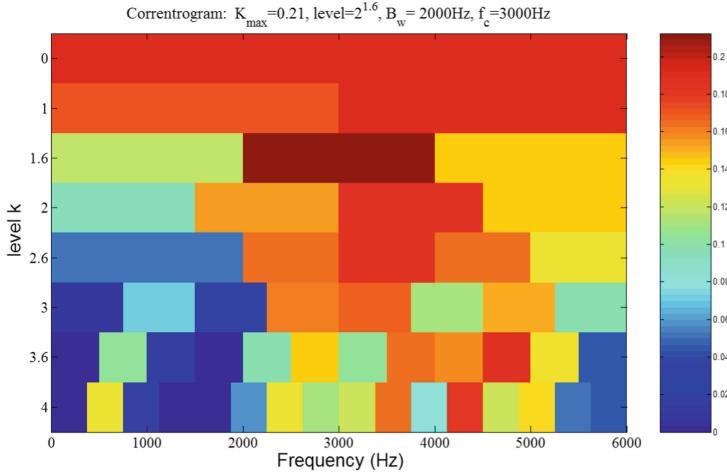


Fig. 9. Correntrogram of bearing inner fault signal

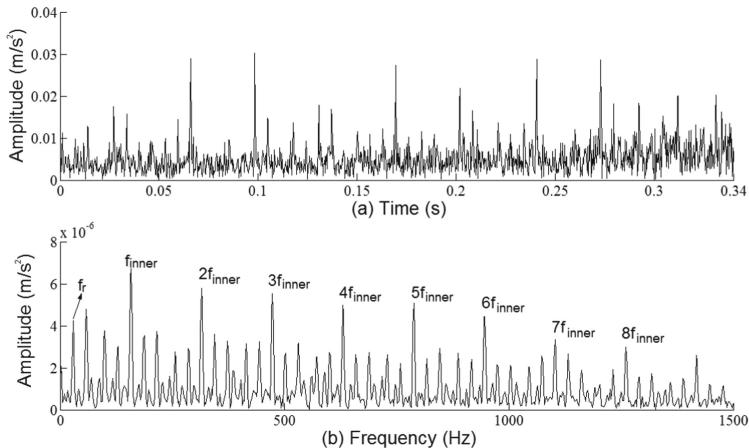


Fig. 10. (a) Envelope of the filtered bearing inner fault signal with maximized Correntrogram; (b) Squared envelope spectrum

For comparison, the traditional Kurtogram is applied to bearing inner fault vibration signal displayed in Fig. 7(a) and the corresponding results are shown in Fig. 11. As seen in Fig. 11, Level 4 has the frequency band with the highest spectral kurtosis value, corresponding to its center frequency and bandwidth being 3562.5 Hz and 375 Hz

respectively. The components extracted by the optimal frequency band and its squared envelope spectrum are shown in Fig. 12(a) and Fig. 12(b) respectively. Although the fault characteristic frequency and its second harmonic of the bearing inner race can be observed in Fig. 12(b), its spectral peak is not very significant and the signal-to-noise ratio is lower. Therefore, it can be clearly exhibited that Correntrogram has significant advantages over traditional Kurtogram in this experiment analysis.

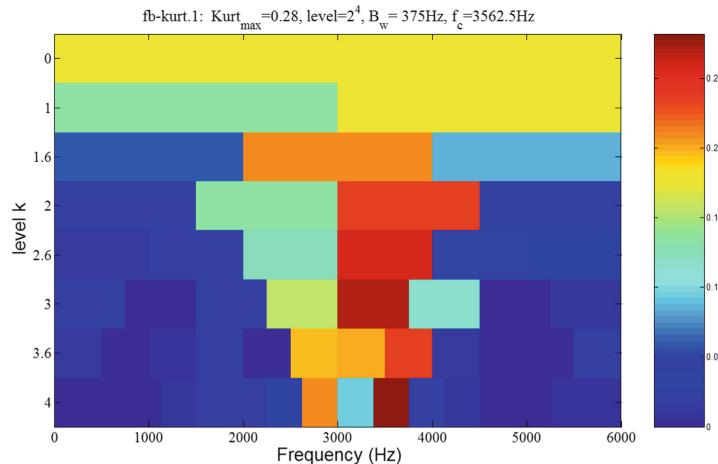


Fig. 11. Kurtogram of bearing inner fault signal

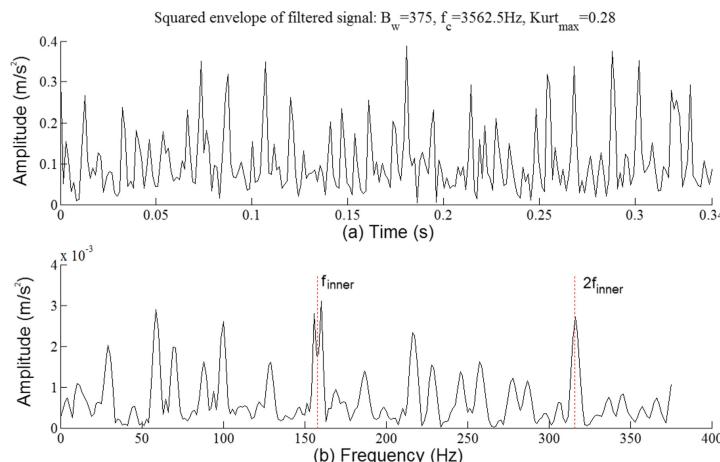


Fig. 12. (a) Amplitude envelope of the filtered bearing inner fault signal with maximized Kurtogram; (b) Squared envelope spectrum

5 Conclusions

A robust optimal frequency band selection method named Correntrogram is put forward. Correntrogram combines the advantages of correntropy and spectral L_2/L_1 norm, which can not only effectively suppress the noise interference in the signal, but also improve the sparsity of the signal. Correntropy is not only an effective tool to deal with Gaussian noise and non-Gaussian noise, but also can effectively enhance the characteristics of weak bearing fault. The superior performance of Correntrogram is confirmed by simulative and bearing inner defect experimental signal analysis.

Acknowledgement. This research is a part of the research that is sponsored by the Wuhu Science and Technology Program (No. 2021jc1–6).

References

1. Rai, A., Upadhyay, S.H.: A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings. *Tribol. Int.* **96**, 289–306 (2016)
2. Abboud, D., Elbadaoui, M., Smith, W.A., et al.: Advanced bearing diagnostics: a comparative study of two powerful approaches. *Mech. Syst. Signal Process.* **114**, 604–627 (2019)
3. Kedadouche, M., Thomas, M., Tahan, A.: A comparative study between empirical wavelet transforms and empirical mode decomposition methods: application to bearing defect diagnosis. *Mech. Syst. Signal Process.* **81**, 88–107 (2016)
4. Antoni, J.: The spectral kurtosis: a useful tool for characterising non-stationary signals. *Mech. Syst. Signal Process.* **20**(2), 282–307 (2006)
5. Antoni, J., Randall, R.B.: The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mech. Syst. Signal Process.* **20**(2), 308–331 (2006)
6. Antoni, J.: Fast computation of the kurtogram for the detection of transient faults. *Mech. Syst. Signal Process.* **21**(1), 108–124 (2007)
7. Xu, T.T., Sun, W.L., Wang, H.W., et al.: Application of variational mode decomposition and spectral kurtosis in gearbox fault diagnosis. *Mech. Sci. Technol. Aeros. Eng.* **39**(6), 873–878 (2020)
8. Antoni, J.: The infogram: entropic evidence of the signature of repetitive transients. *Mech. Syst. Signal Process.* **74**, 73–94 (2016)
9. Santamaría, I., Pokharel, P.P., Principe, J.C.: Generalized correlation function: definition, properties, and application to blind equalization. *IEEE Trans. Signal Process.* **54**(6), 2187–2197 (2006)
10. Liu, W., Pokharel, P.P., Principe, J.C.: Correntropy: properties and applications in non-gaussian signal processing. *IEEE Trans. Signal Process.* **55**(11), 5286–5298 (2007)
11. Li, H., Hao, R.: Rolling bearing fault diagnosis based on sensor information fusion and generalized cyclic cross correntropy spectrum density. *J. Vib. Shock* **41**(2), 200–207 (2022)



Semidefinite Relaxation Algorithm for Source Localization Using Multiple Groups of TDOA Measurements with Distance Constraints

Tao Zhang, Wuyi Yang^(✉), and Yu Zhang^{ID}

Key Laboratory of Underwater Acoustic Communication and Marine Information Technology
of the Ministry of Education, College of Ocean and Earth Sciences, Xiamen University,
Xiamen 361000, People's Republic of China
wyyang@xmu.edu.cn

Abstract. A semidefinite relaxation (SDR) algorithm is proposed for scenarios in which sources, such as locator beacons and marine mammals, transmit signals continuously at multiple locations. In these scenarios, multiple groups of the time-difference-of-arrival (TDOA) measurements can be obtained, and each group of the TDOA measurements corresponds to one source location. The proposed algorithm adds constraints on the distances between the source locations and jointly locates the positions of the source using multiple groups of the TDOA measurements. The objective function of the maximum likelihood (ML) localization estimation is nonconvex, and the SDR is adopted to approximate the nonconvex ML optimization by relaxing it to a semidefinite programming (SDP) problem. The simulation results show the superior performance of the proposed method.

Keywords: Source localization · Time difference of arrival measurements · Semidefinite relaxation · Semidefinite programming

1 Introduction

Source localization is of great significance for a variety of applications such as navigation, target tracking, emergency response and others. A common approach is to use the time-difference-of-arrival (TDOA) measurements, that is, the differences in arrival times between pairs of sensor outputs which receive the emitted signal. At least three sensors are required to uniquely estimate the position of the source in a two-dimensional (2D) plane, and three-dimensional (3D) positioning requires four or more sensors.

Source localization using the hyperbolic equations constructed from the TDOA measurements is a nonlinear and nonconvex problem. Many researchers have made significant exploration to tackle this problem. The well-known two-step weighted least squares (WLS) method [1] reorganizes the nonlinear equations into a set of linear equations by squaring them and introducing an extra variable that is a function of the source position. Then, the relationship between the extra variable and the source position is utilized to achieve better estimation. Based on the nonlinear least-squares (NLS) framework, the Taylor-series method [2] utilizes Taylor-series expansion for linearizing the nonlinear

equations and then performs gradient searches for the minimum in an iterative manner. This method suffers from initial condition sensitivity and convergence difficulty. Starting from the maximum likelihood (ML) function, the approximate ML method [3] derived a closed-form approximate solution to the ML equations.

Using convex optimization methods to solve the localization problems has attracted great interests. Among various convex optimization techniques, semidefinite programming (SDP) can be solved globally in polynomial time and provides impressive modeling capability [4]. Several kinds of SDP approaches [5–8] were devised to approximating the location problems in TDOA localization. Wang and Ansari [5] reformulated the localization problem based on the WLS criterion and then solved it by semidefinite relaxation (SDR). This method requires an initial estimate to formulate the optimization problem and if the starting point is not accurate enough, the performance degradation occurs. Wang and Wu [6] reformulated the localization problem based on the robust LS criterion and then performed SDR to obtain a convex semidefinite programming (SDP) problem, which can be solved efficiently via optimization toolbox. This method does require the initial estimate. Zou *et al.* [7] formulated the source localization problem using TDOA measurements that are collected under non-line-of-sight (NLOS) conditions as a robust least squares (RLS) problem and propose two efficiently implementable convex relaxation-based approximation methods to the RLS problem. Zou *et al.* [8] developed a maximum likelihood estimator (MLE) for joint synchronization and localization method in wireless sensor networks and transformed the nonconvex MLE problem into a convex optimization problem.

In some scenarios, sources, such as locator beacons (LBs) and marine mammals, transmit signals continuously at multiple locations. Underwater LBs transmit short-duration pulses for localization, which can be used to monitor divers, search and rescue operations, etc. [9]. Marine mammals such as sperm whales, beaked whales and delphinids produce short duration, broadband echolocation clicks quite frequently for navigation, prey detection, and communication [10]. Based on passive acoustic monitoring (PAM), marine mammals can be detected [11], identified [12], and located [13, 14]. In these scenarios, multiple groups of the TDOA measurements of a source can be obtained, in which each group of the TDOA measurements corresponds to one source location. Assuming that the max velocity of the source is known, the distance constraints between the source locations can be derived based on the time of arrival (TOA) measurements of the source. A semidefinite relaxation algorithm is proposed which adds the distance constraints and jointly locates the positions of the source using multiple groups of the TDOA measurements. The SDR is used to approximate the nonconvex maximum likelihood (ML) optimization by relaxing it to an SDP problem. Simulations are performed to evaluate the effectiveness of the proposed method.

2 Proposed Method

2.1 Problem Formulation

A m -dimensional (with $m = 2$ or 3) scenario is considered, and a source emits signals at locations $\mathbf{u}_j, j = 1, \dots, M$. Consider a network composed of N sensors which receive these signals. When the j -th signal is received, locations of these sensors are known and

denoted by $\mathbf{s}_{i,j}$, $i = 1, \dots, N, j = 1, \dots, M$, respectively. With sensors synchronized to a common clock, the j -th signal arrives at the i -th sensor with TOA value $t_{i,j}$

$$t_{i,j} = \frac{1}{c_s} \|\mathbf{u}_j - \mathbf{s}_{i,j}\| + t_{0,j} + \xi_{i,j}, i = 1, \dots, N, j = 1, \dots, M, \quad (1)$$

where c_s is the signal propagation speed, $t_{0,j}$ is the unknown reference time at which the signal was transmitted, and $\xi_{i,j}$ denotes the TOA measurement noise that is assumed to be zero-mean Gaussian distributed. The first hydrophone is chosen as the reference hydrophone. The TDOA measurement between a sensor pair i and 1 is given by

$$\tau_{i,j} = \frac{1}{c_s} \|\mathbf{u}_j - \mathbf{s}_{i,j}\| - \frac{1}{c_s} \|\mathbf{u}_j - \mathbf{s}_{1,j}\| + \xi_{i,j} - \xi_{1,j}, i = 2, \dots, N, j = 1, \dots, M. \quad (2)$$

The distance between the source and the i -th sensor is $r_{i,j} = \|\mathbf{u}_j - \mathbf{s}_{i,j}\|, i = 1, \dots, N, j = 1, \dots, M$, and $\|\cdot\|$ is the Euclid norm. Multiplying c_s to TDOA measurements in (2), the range difference of arrival (RDOA) is obtained by

$$d_{i,j} = \|\mathbf{u}_j - \mathbf{s}_{i,j}\| - \|\mathbf{u}_j - \mathbf{s}_{1,j}\| + n_{i,j}, i = 2, \dots, N, j = 1, \dots, M, \quad (3)$$

where $n_{i,j} = c_s(\xi_{i,j} - \xi_{1,j})$. Then we can get the following vector form

$$\mathbf{d}_j = \mathbf{G}\mathbf{r}_j + \mathbf{n}_j, j = 1, \dots, M, \quad (4)$$

where $\mathbf{G} = [-\mathbf{1}_{(N-1) \times 1}, \mathbf{I}_{(N-1) \times (N-1)}]$, $\mathbf{d}_j = [d_{2,j}, \dots, d_{N,j}]^T$, $\mathbf{r}_j = [r_{1,j}, r_{2,j}, \dots, r_{N,j}]^T$, and $\mathbf{n}_j = [n_{2,j}, \dots, n_{N,j}]^T$. $\mathbf{1}_{(N-1) \times 1}$ is the $(N-1) \times 1$ all one column vector, and $\mathbf{I}_{(N-1) \times (N-1)}$ is the $(N-1) \times (N-1)$ identity matrix.

Assuming that the max velocity v_{\max} of the source is known, the distance constraints between the source locations can be derived based on the TOA measurements of the source. The distance between \mathbf{u}_j and \mathbf{u}_{j+1} can be bounded by a certain value $\delta_{j,j+1} = v_{\max}(t_{1,j+1} - t_{1,j})$ i.e., $\|\mathbf{u}_j - \mathbf{u}_{j+1}\| \leq \delta_{j,j+1}, j = 1, \dots, M-1$.

2.2 Maximum Likelihood Estimation of Source Positions

Under the assumption that the TOA measurements are Gaussian distributed, the joint conditional probability density of $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_M]$ is

$$p(\mathbf{D}|\mathbf{U}) = (2\pi\sigma^2)^{-\frac{(N-1)M}{2}} |\mathbf{Q}|^{-\frac{M}{2}} \exp \left\{ \sum_{j=1}^M -\frac{(\mathbf{d}_j - \mathbf{G}\mathbf{r}_j)^T \mathbf{Q}^{-1} (\mathbf{d}_j - \mathbf{G}\mathbf{r}_j)}{2\sigma^2} \right\}, \quad (5)$$

where σ^2 is the range measurement error variance, $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_M]$, and $\mathbf{Q} = \begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{bmatrix}$.

Incorporating the distance constraints $\|\mathbf{u}_j - \mathbf{u}_{j+1}\| \leq \delta_{j,j+1}, j = 1, \dots, M-1$, the ML estimation of \mathbf{U} can be written as a constrained quadratic program as follows:

$$\begin{aligned} & \min_{\mathbf{U}} \sum_{j=1}^M (\mathbf{d}_j - \mathbf{G}\mathbf{r}_j)^T \mathbf{W}(\mathbf{d}_j - \mathbf{G}\mathbf{r}_j), \\ \text{s.t. } & [\mathbf{r}_j]_i^2 = r_{i,j}^2 = \|\mathbf{u}_j - \mathbf{s}_{i,j}\|^2, i = 1, 2, \dots, N, j = 1, \dots, M, \\ & \|\mathbf{u}_j - \mathbf{u}_{j+1}\| \leq \delta_{j,j+1}, j = 1, \dots, M-1, \end{aligned} \quad (6)$$

where $\mathbf{W} = \mathbf{Q}^{-1}$. It is challenging to solve this cost function which is highly nonlinear and nonconvex.

To transform the source localization problem to a standard convex optimization problem, the formulas in (6) can be reformulated as

$$(\mathbf{d}_j - \mathbf{G}\mathbf{r}_j)^T \mathbf{W}(\mathbf{d}_j - \mathbf{G}\mathbf{r}_j) = \text{tr} \left\{ \begin{bmatrix} \mathbf{R}_j & \mathbf{r}_j \\ \mathbf{r}_j^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{G}^T \mathbf{W} \mathbf{G} & -\mathbf{G}^T \mathbf{W} \mathbf{d}_j \\ -\mathbf{d}_j^T \mathbf{W} \mathbf{G} & \mathbf{d}_j^T \mathbf{W} \mathbf{d}_j \end{bmatrix} \right\}, \quad (7)$$

$$[\mathbf{r}_j]_i^2 = r_{i,j}^2 = \text{Tr} \left(\begin{bmatrix} \mathbf{I} & \mathbf{u}_j \\ \mathbf{u}_j^T & [\mathbf{B}]_{jj} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{i,j} \mathbf{s}_{i,j}^T & -\mathbf{s}_{i,j} \\ -\mathbf{s}_{i,j}^T & 1 \end{bmatrix} \right), \quad (8)$$

$$\|\mathbf{u}_j - \mathbf{u}_{j+1}^2\| = \text{tr} \left\{ \mathbf{B} \mathbf{E}^{j,j+1} \right\}, \quad (9)$$

where $\text{tr}(\cdot)$ is the trace of an input matrix, $\mathbf{R}_j = \mathbf{r}_j \mathbf{r}_j^T$, $\mathbf{B} = \mathbf{U}^T \mathbf{U}$, $\mathbf{E}^{j,j+1} = (\mathbf{e}_j - \mathbf{e}_{j+1})(\mathbf{e}_j - \mathbf{e}_{j+1})^T$, and $\mathbf{e}_j = [\mathbf{0}_{1 \times (j-1)}, \mathbf{1}, \mathbf{0}_{1 \times (M-j)}]^T$.

In order to transform the problem in (6) to a standard convex optimization problem, we relax the constraints $\mathbf{R}_j = \mathbf{r}_j \mathbf{r}_j^T$ as $\mathbf{R}_j \succeq \mathbf{r}_j \mathbf{r}_j^T, j = 1, \dots, M$, and $\mathbf{B} = \mathbf{U}^T \mathbf{U}$ as $\mathbf{B} \succeq \mathbf{U}^T \mathbf{U}$, respectively. Additionally, $[\mathbf{R}_j]_{ii}$ can be expressed as

$$\begin{aligned} [\mathbf{R}_j]_{ii} &= \text{Tr} \left(\begin{bmatrix} \mathbf{I} & \mathbf{u}_j \\ \mathbf{u}_j^T & [\mathbf{B}]_{jj} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{i,j} \mathbf{s}_{i,j}^T & -\mathbf{s}_{i,j} \\ -\mathbf{s}_{i,j}^T & 1 \end{bmatrix} \right), \\ & i = 1, 2, \dots, N, j = 1, \dots, M. \end{aligned} \quad (10)$$

To yield precise solution, additional constraints for $\mathbf{R}_j, j = 1, \dots, M$ should be added to enhance the tightness. By Cauchy-Schwartz inequality, $r_{i,j} r_{k,j} = \|\mathbf{u}_j - \mathbf{s}_{i,j}\| \|\mathbf{u}_j - \mathbf{s}_{k,j}\| \geq |(\mathbf{u}_j - \mathbf{s}_{i,j})^T (\mathbf{u}_j - \mathbf{s}_{k,j})|$, the following inequalities can be derived:

$$\begin{aligned} [\mathbf{R}_j]_{ik} &\geq \left| \text{Tr} \left(\begin{bmatrix} \mathbf{I} & \mathbf{u}_j \\ \mathbf{u}_j^T & [\mathbf{B}]_{jj} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{i,j} \mathbf{s}_{k,j}^T & -\mathbf{s}_{k,j} \\ -\mathbf{s}_{i,j}^T & 1 \end{bmatrix} \right) \right|, \\ & j = 1, \dots, M, i, k = 1, \dots, N, k > i. \end{aligned} \quad (11)$$

Based on SDR, the formulas (7), (8), and (9), and the constraints (10) and (11), the convex SDP problem is obtained as follows:

$$\min_{\mathbf{U}, \mathbf{B}, \mathbf{r}_1, \dots, \mathbf{r}_M, \mathbf{R}_1, \dots, \mathbf{R}_M} \sum_{j=1}^M \text{tr} \left\{ \begin{bmatrix} \mathbf{R}_j & \mathbf{r}_j \\ \mathbf{r}_j^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{G}^T \mathbf{W} \mathbf{G} & -\mathbf{G}^T \mathbf{W} \mathbf{d}_j \\ -\mathbf{d}_j^T \mathbf{W} \mathbf{G} & \mathbf{d}_j^T \mathbf{W} \mathbf{d}_j \end{bmatrix} \right\} \quad (12)$$

s.t.

$$\begin{aligned} & \begin{bmatrix} \mathbf{B} & \mathbf{U}^T \\ \mathbf{U} & \mathbf{I} \end{bmatrix} \succeq \mathbf{0}, \\ & \begin{bmatrix} \mathbf{R}_j & \mathbf{r}_j \\ \mathbf{r}_j^T & \mathbf{I} \end{bmatrix} \succeq \mathbf{0}, j = 1, \dots, M, \\ & \text{tr}\{\mathbf{B}\mathbf{E}^{j+1}\} \leq \delta_{j,j+1}^2, j = 1, \dots, M-1, \\ [\mathbf{R}_j]_{ii} &= \text{Tr}\left(\begin{bmatrix} \mathbf{I} & \mathbf{u}_j \\ \mathbf{u}_j^T & [\mathbf{B}]_{jj} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{i,j} \mathbf{s}_{i,j}^T & \mathbf{s}_{i,j} \\ -\mathbf{s}_{i,j}^T & 1 \end{bmatrix}\right), j = 1, \dots, M, i = 1, \dots, N, \\ [\mathbf{R}_j]_{ik} &\geq \left| \text{Tr}\left(\begin{bmatrix} \mathbf{I} & \mathbf{u}_j \\ \mathbf{u}_j^T & [\mathbf{B}]_{jj} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{i,j} \mathbf{s}_{k,j}^T & -\mathbf{s}_{k,j} \\ -\mathbf{s}_{i,j}^T & 1 \end{bmatrix}\right) \right|, \\ & j = 1, \dots, M, i, k = 1, \dots, N, k > i \end{aligned}$$

By solving the SDP in (12), the estimation of \mathbf{U} is finally obtained.

2.3 Complexity Analysis

Given a solution accuracy $\varepsilon > 0$, the worst-case complexity of solving SDP is $O\left(\max\{k_1, k_2\}^4 k_2^{\frac{1}{2}} \log(1/\varepsilon)\right)$, where k_1 is the number of constraints and k_2 is the problem size [15]. In the SDP (12), we have $k_1 = (N(N+1)/2 + 1)M - 1$ and $k_2 = (N+1)M$. Hence, the complexity of (12) is roughly $O\left((MN(N+1)/2)^4 (NM)^{\frac{1}{2}} \log(1/\varepsilon)\right)$.

2.4 Constrained Cramér-Rao Lower Bound

The Cramér-Rao inequality [1] sets a lower bound for the variance of any unbiased parameter estimators. The Fisher information matrix (FIM) is obtained as

$$\mathbf{F} = -E\left[\frac{\partial^2 \ln(p(\mathbf{D}|\mathbf{z}))}{\partial \mathbf{z} \partial \mathbf{z}^T}\right] = \frac{1}{\sigma^2} \text{blkdiag}\{\mathbf{F}_1, \dots, \mathbf{F}_M\} \quad (13)$$

where $\mathbf{z} = [\mathbf{u}_1^T, \dots, \mathbf{u}_M^T]^T$, $p(\mathbf{D}|\mathbf{z}) = p(\mathbf{D}|\mathbf{U})$, $\text{blkdiag}(\mathbf{F}_1, \dots, \mathbf{F}_M)$ returns the block diagonal matrix created by aligning the input matrices $\mathbf{F}_1, \dots, \mathbf{F}_M$ along the diagonal of \mathbf{F} .

$$\mathbf{F}_j = \left[\mathbf{G} \frac{\partial \mathbf{r}_j}{\partial \mathbf{u}_j^T} \right]^T \mathbf{W} \left[\mathbf{G} \frac{\partial \mathbf{r}_j}{\partial \mathbf{u}_j^T} \right], j = 1, \dots, M, \quad (14)$$

and

$$\frac{\partial \mathbf{r}_j}{\partial \mathbf{u}_j^T} = \left[\frac{\mathbf{u}_j - \mathbf{s}_{1,j}}{\|\mathbf{u}_j - \mathbf{s}_{1,j}\|}, \dots, \frac{\mathbf{u}_j - \mathbf{s}_{N,j}}{\|\mathbf{u}_j - \mathbf{s}_{N,j}\|} \right]^T, j = 1, \dots, M. \quad (15)$$

For any unbiased estimate $\hat{\mathbf{z}}$, the Cramér-Rao Lower Bound (CRLB) is $E(\|\hat{\mathbf{z}} - \mathbf{z}\|^2) \geq \text{Trace}(\mathbf{F}^{-1})$. Based on the *Theorem 1* in [11], the constrained CRLB (CCRLB) can be derived. The inequalities in (8) can be written as

$$\mathbf{c}(\mathbf{z}) = [c_{1,2}(\mathbf{z}), \dots, c_{M-1,M}(\mathbf{z})]^T \leq \mathbf{0}_{(M-1) \times 1}, \quad (16)$$

where $c_{j,j+1}(\mathbf{z}) = \|\mathbf{u}_j - \mathbf{u}_{j+1}\| - \delta_{j,j+1}$, $j = 1, \dots, M-1$. For any unbiased estimator \mathbf{z} , the estimator error covariance matrix satisfies the matrix inequality [7]

$$\Sigma_{\mathbf{z}} \succeq \mathbf{H}_{\mathbf{z}} = \mathbf{Q}_{\mathbf{z}} \mathbf{F}^{-1}, \quad (17)$$

where $\mathbf{Q}_{\mathbf{z}}$ is defined as follows:

$$\mathbf{Q}_{\mathbf{z}} = \mathbf{I} - \mathbf{F}^{-1} \left[\frac{\partial \mathbf{c}(\mathbf{z})}{\partial \mathbf{z}^T} \right]^T \left\{ \frac{\partial \mathbf{c}(\mathbf{z})}{\partial \mathbf{z}^T} \mathbf{F}^{-1} \left[\frac{\partial \mathbf{c}(\mathbf{z})}{\partial \mathbf{z}^T} \right]^T \right\}^+ \frac{\partial \mathbf{c}(\mathbf{z})}{\partial \mathbf{z}^T}, \quad (18)$$

where the $\{\cdot\}^+$ denotes pseudo-inverse,

$$\frac{\partial \mathbf{c}(\mathbf{z})}{\partial \mathbf{z}^T} = \begin{bmatrix} \frac{\partial c_{1,2}(\mathbf{z})}{\partial \mathbf{z}^T} \\ \vdots \\ \frac{\partial c_{M-1,M}(\mathbf{z})}{\partial \mathbf{z}^T} \end{bmatrix}, \quad (19)$$

and

$$\frac{\partial c_{j,j+1}(\mathbf{z})}{\partial \mathbf{z}^T} = \left[\mathbf{0}_{1 \times 3(j-1)} \frac{(\mathbf{u}_j - \mathbf{u}_{j+1})^T}{\|\mathbf{u}_j - \mathbf{u}_{j+1}\|} - \frac{(\mathbf{u}_j - \mathbf{u}_{j+1})^T}{\|\mathbf{u}_j - \mathbf{u}_{j+1}\|} \mathbf{0}_{1 \times 3(M-j-1)} \right], \quad (20)$$

$$j = 1, \dots, M-1.$$

Incorporating the distance constraint inequalities, the CCRLB is $(\|\hat{\mathbf{z}} - \mathbf{z}\|^2) \geq \text{Trace}(\mathbf{H}_{\mathbf{z}})$, for any unbiased estimate $\hat{\mathbf{z}}$. The constrained parameter space Θ_C is defined by $\mathbf{c}(\mathbf{z})$. When \mathbf{z} lies in the interior of Θ_C , namely, \mathbf{z} belongs to the set $\{\mathbf{z} : \mathbf{c}(\mathbf{z}) < \mathbf{0}\}$ where all the equality constraints are inactive, the CCRLB is identical to the CRLB. When all the equality constraints are active, i.e., $\mathbf{c}(\mathbf{z}) = \mathbf{0}$, the CCRLB is identical to the CRLB with the equality constraints, thus leading to the most bound reduction.

3 Experiment and Analysis

The proposed algorithm is denoted as SDP-M, while labelling the algorithms in [1] and [6] as TS-WLS and SDPE, respectively. Simulations were conducted to evaluate the performance of SDP-M with TS-WLS, SDP-E, CRLB as well as CCRLB. A 3D underwater scenario was considered, and a sound source emitted a series of pulses. It was assumed that 4 hydrophones were stationary in seawater and their positions are listed in Table 1, in which z represents the hydrophone depth. Given the short propagation path from the sound source to the hydrophones, the sound speed variation with depth and range was assumed to cause negligible refraction [10]. Considering multipath propagation, the

TDOA measurements of the direct pulses and the reflected pulses by the sea surface were used for positioning [9, 10]. The positioning performance was evaluated by the root mean square errors (RMSEs) of the source positions. The value of σ^2 was modified to achieve different noise conditions, and all results were averages of 200 independent runs. The SDP based methods were solved using MATLAB toolbox CVX [17].

Table 1. Positions of hydrophones (m).

Hydrophone no.	1	2	3	4
x	0	0	50	50
y	0	50	0	50
z	25	50	75	50

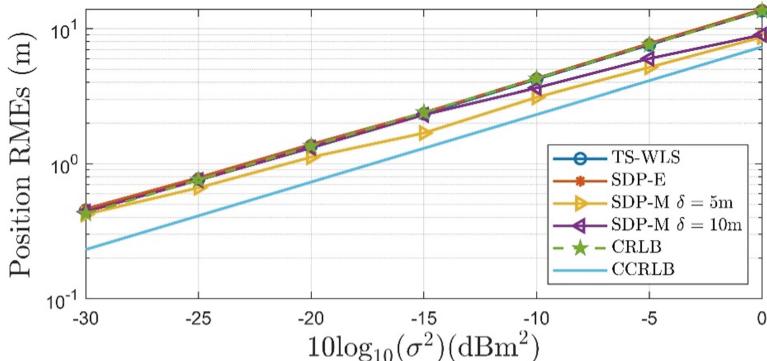


Fig. 1. Comparison of position RMSEs against the measurement error variance using different methods in the near-field scenario.

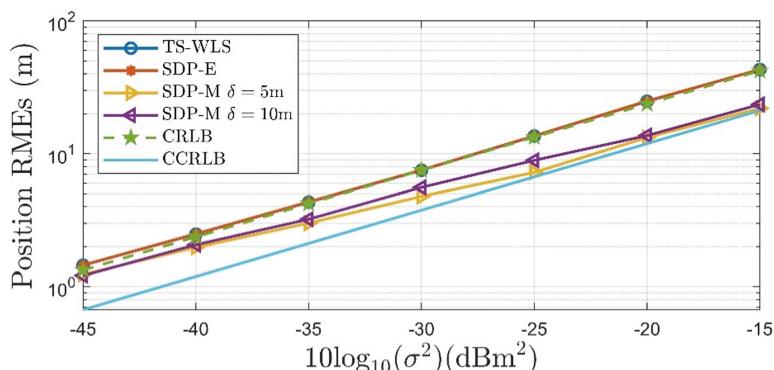


Fig. 2. Comparison of position RMSEs against the measurement error variance using different methods in the far-field scenario.

The near-field localization scenario was first considered. Assuming that the sound source was initially located at (100 m, 150 m, 100 m), the sound source then moved to (103 m, 152 m, 101 m), (106 m, 154 m, 102 m), (109 m, 156 m, 103 m), and emitted a pulse at each position respectively. Thus, the distance between the adjacent source positions was 3.742 m, and two distance bounds $\delta_{j,j+1} = 5$ m and $\delta_{j,j+1} = 10$ m, $j = 1, 2, 3$, were considered respectively. The estimated position RMSEs are shown in Fig. 1. Then, the far-field scenario was considered. Assuming that the sound source was initially located at (500 m, 550 m, 310 m), the sound source then moved to (503 m, 552 m, 311 m), (506 m, 554 m, 312 m), (509 m, 556 m, 313 m). The experimental results are shown in Fig. 2. It is obvious that the proposed method performed best.

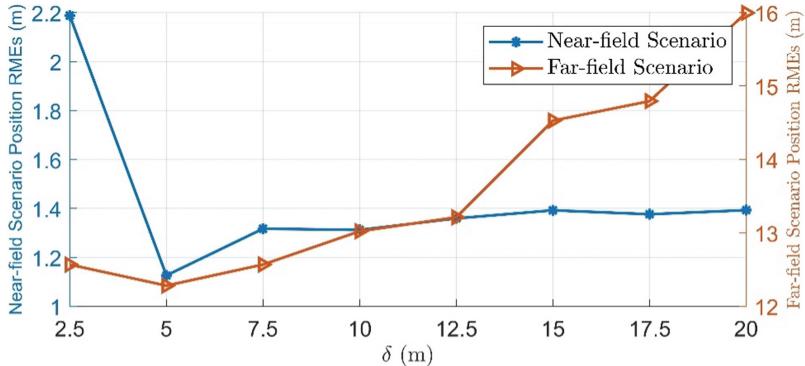


Fig. 3. Position RMSEs with respect to different values of the distance bound when $10 \log_{10}(\sigma^2) = -20$ dBm².

As shown in Fig. 1 and Fig. 2, the distance constraints were more effective as the measurement noise increased. Figure 3 shows the estimated position RMSEs with respect to different values of the distance bound δ when $10 \log_{10}(\sigma^2) = -20$ dBm². When the distance bound was more precise, the proposed method achieved better results.

Experiments were performed on a desktop computer with an Intel i7-8565U CPU and 16 GB of RAM. The average CPU calculation time of estimating the four locations was 1.3 ms, 2070.1 ms, and 2783.7 ms for TS-WLS, SDP-E and SDP-M, respectively. SDP-M is computationally higher than TS-WLS and SDP-E.

4 Conclusion

For locating a source which transmits signals at multiple locations, an algorithm has been proposed. To improve localization accuracy, the proposed algorithm adds constraints on the distances between the source locations and jointly locates the positions of the source using multiple groups of the TDOA measurements, in which each group of the TDOA measurements corresponds to one source location. The SDR was adopted to approximate the nonconvex ML localization estimation by relaxing it to an SDP problem. The simulation results show the superior performance of the proposed method, and the distance constraints are effective, especially when the measurement noise is large.

References

1. Chan, Y.T., Ho, K.C.: A simple and efficient estimator for hyperbolic location. *IEEE Trans. Sig. Process.* **42**(8), 1905–1915 (1994)
2. Spirito, M.A.: On the accuracy of cellular mobile station location estimation. *IEEE Trans. Veh. Technol.* **50**(3), 674–685 (2001)
3. Chan, Y., Hang, H., Ching, P.: Exact and approximate maximum likelihood localization algorithms. *IEEE Trans. Veh. Technol.* **55**(1), 10–16 (2006)
4. Nesterov, Y.: *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-91578-4>
5. Wang, G., Li, Y., Ansari, N.: A semidefinite relaxation method for source localization using TDOA and FDOA measurements. *IEEE Trans. Veh. Technol.* **62**(2), 853–862 (2013)
6. Wang, Y., Wu, Y.: An efficient semidefinite relaxation algorithm for moving source localization using TDOA and FDOA measurements. *IEEE Commun. Lett.* **21**(1), 80–83 (2017)
7. Wang, G., So, A.M., Li, Y.: Robust convex approximation methods for TDOA-based localization under NLOS conditions. *IEEE Trans. Sig. Process.* **64**(13), 3281–3296 (2016)
8. Zou, Y., Liu, H., Wan, Q.: Joint synchronization and localization in wireless sensor networks using semidefinite programming. *IEEE Internet Things J.* **5**(1), 199–205 (2018)
9. Skarsoulis, E.K., Dosso, S.E.: Linearized two-hydrophone localization of a pulsed acoustic source in the presence of refraction: theory and simulations. *J. Acoust. Soc. Am.* **138**(4), 2221–2234 (2015)
10. Tran, D.D., et al.: Using a coherent hydrophone array for observing sperm whale range, classification, and shallow-water dive profiles. *J. Acoust. Soc. Am.* **135**(6), 3352–3363 (2014)
11. Luo, W., Yang, W., Zhang, Y.: Convolutional neural network for detecting odontocete echolocation clicks. *J. Acoust. Soc. Am.* **145**(1), EL7–EL12 (2019)
12. Yang, W., Luo, W., Zhang, Y.: Classification of odontocete echolocation clicks using convolutional neural network. *J. Acoust. Soc. Am.* **147**(1), 49–55 (2020)
13. Rideout, B.P., Dosso, S.E., Hannay, D.E.: Underwater passive acoustic localization of Pacific walruses in the northeastern Chukchi Sea. *J. Acoust. Soc. Am.* **134**(3), 2534–2545 (2013)
14. Barlow, J., Griffiths, E.T., Klinck, H., Harris, D.V.: Diving behavior of Cuvier’s beaked whales inferred from three-dimensional acoustic localization and tracking using a nested array of drifting hydrophone recorders. *J. Acoust. Soc. Am.* **144**(4), 2030–2041 (2018)
15. Gorman, J., Hero, A.: Lower bounds for parametric estimation with constraints. *IEEE Trans. Inform. Theory* **26**(6), 1285–1301 (1990)
16. Luo, Z.Q., Ma, W.K., So, A.M.C., Ye, Y., Zhang, S.: Semidefinite relaxation of quadratic optimization problems. *IEEE Sig. Process. Mag.* **27**(3), 20–34 (2010)
17. CVX Research: CVX: Matlab software for disciplined convex programming, version 2.2. <http://cvxr.com/cvx>. Accessed 31 Jan 2020

Pattern Recognition



Quasi Fourier Descriptor for Affine Invariant Features

Chengyun Yang, Lei Lu^(✉), Lei Zhang, Yu Tian, and Zhang Chen

School of Mechanical and Electrical Engineering, Soochow University, Suzhou 215000, China
lulei@suda.edu.cn

Abstract. The extraction of affine invariant features plays an important role in many computer vision tasks. Region-based methods can achieve high accuracy with expensive computation. Whereas, contour-based techniques need less computation but their performance is strongly dependant on the boundary extraction. To combine region-based and contour-based methods together for data reduction, *central radial transform* (CRT) is proposed in this paper. Any object is converted into a closed curve. Then, Fourier descriptor is conducted on the obtained closed curve. The derived features are invariant to affine transform, and are called *quasi Fourier descriptors* (QFDs). In compare with some related methods, CRT has eliminated shearing in affine transform. Consequently, parameterization and extra transforms have been avoided. Experimental results show affine invariance of the proposed QFD. In comparison with some region-based methods, the proposed QFD is more robust to additive noise.

Keywords: Quasi Fourier descriptors (QFDs) · Fourier descriptor (FD) · Affine transform · Affine invariant feature

1 Introduction

For an object, if its images are derived from different viewpoints, these images often undergo geometric deformations. These geometric deformations can be approximated with affine transform [1–3]. As a result, the extraction of affine invariant features plays an important role in many object recognition tasks, and has been found many applications ([4, 5], etc.). Various techniques have been developed for the extraction of affine invariant features ([1, 6, 7], etc.). Based on whether features are extracted from the contour only or from the whole shape region, the existing methods can be classified into two categories [8]: contour-based methods and region-based methods.

Fourier descriptor (FD) [9] has become the most widely utilized contour-based technique. It is based on the well-developed Fourier transform, and can achieve attractive invariance properties. Consequently, many contour-based methods related to FD have been proposed ([10–13], etc.). Arbter et al. [3] have modified the traditional FD such that it can be utilized to extract affine invariant features. Contour-based methods are often of better data reduction, but they are strongly dependant on the extraction of contours. These methods are usually only applicable to objects with single boundary. They are

invalid to objects which consist of several components. Consequently, contour-based methods are limited to some applications.

In contrast to contour-based methods, region-based methods usually achieve high accuracy, but some of them are of high computational demands. *Affine moment invariants* (AMIs) [1, 14] are the most famous region-based method for affine invariant features. To improve the robustness of moment-based methods to noise, Yang [15] propose *cross-weighted moment* (CWM). But the computational cost of CWM is extremely expensive. Afterwards, *multi-scale autoconvolution* (MSA) has been put forward by Rahtu et al. [16]. However, the computational complexity of MSA is still large.

Contour-based and region-based techniques both have their merits and short comes, thus an intuitive way is to combine them together. In this paper, *central radial transform* (CRT) is proposed to convert any image into a closed curve. All pixels in shape region have been utilized. Then FD is conducted on the obtained closed curve. The used technique is contour-based. The derived features are called *quasi Fourier descriptors* (QFDs). They are invariant to affine transform. Experimental results show the affine invariance of the proposed QFD. In comparison with AMIs and MSA, QFD is more robust to additive noise.

With QFDs, parameterization and some extra transformations have been avoided. Recently, several approaches have been proposed to combine region-based and contour-based methods together for saving the computational cost. In [17], central projection has been employed to convert an object into a closed curve. Then parameterization and whitening transform are conducted on the derived closed curve. Finally, FD is applied to extract affine invariant features. In [18], *polar radius integral transform* (PRIT) is proposed to convert an object into a closed curve. Then parameterization and stationary wavelet transform are conducted on the obtained closed curve for affine invariant features. In comparison with these methods, the proposed QFD is obtained by directly applying FD on the derived closed curve. Parameterization and whitening transform (or wavelet transform) are avoided.

The rest of this paper is organized as follows: In Sect. 2, some basic materials are presented. QFD is proposed in Sect. 3. Properties of QFD have also been discussed. Experimental results are shown in Sect. 4. Finally, some conclusion remarks are provided in Sect. 5.

2 Preliminary

In this section, some background materials are provided. The concept of affine transform is presented, and the traditional FD is reviewed.

2.1 Affine Transform

Affine transform is the transformation defined as follows [1]

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}, \quad (1)$$

where $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a non-singular matrix. Affine transform is the best linear approximation model for perspective distortions [1, 2]. It not only includes similarity transform (translation, rotation and scaling), but also includes shearing. In order to achieve translation invariance, the origin is translated to $O(x_0, y_0)$, the centroid of image $I(x, y)$. Here,

$$x_0 = \frac{\iint xI(x, y)dx dy}{\iint I(x, y)dx dy}, \quad y_0 = \frac{\iint yI(x, y)dx dy}{\iint I(x, y)dx dy}. \quad (2)$$

Consequently, Eq. (1) is converted into the following relation

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix}. \quad (3)$$

2.2 Traditional FD

FD is obtained by applying Fourier transform on a shape signature. A shape signature is a one-dimensional function acquired from the object (often from the boundary). Shape signature can either be real or complex. Many different FDs have been proposed with diverse shape signatures. The performance of shape signatures directly affects the performance of the derived FDs.

Suppose that $s(\theta)$ ($\theta = 0, 1, 2, \dots, N - 1$) is the shape signature. Discrete Fourier transform is conducted on $s(\theta)$ as follows

$$S_k = \frac{1}{N} \sum_{\theta=0}^{N-1} s(\theta) e^{-i \frac{2\theta\pi k}{N}}, \quad k = 0, 1, 2, \dots, N - 1. \quad (4)$$

Consequently, $\{S_k\}$ in Eq. (4) is employed to construct the following FDs

$$C_k = \frac{|S_{k+2}|}{|S_1|}, \quad k = 0, 1, 2, \dots, N - 3. \quad (5)$$

The traditional FD is only invariant to translation, rotation, and scaling (similarity transform). As mentioned previously, similarity transform is only a special case of affine transform. We will modify the traditional FD such that it can be used to extract affine invariant features.

3 Affine Invariant Features with QFD

The definition of QFD is provided. Then properties of QFD are discussed.

3.1 Definition of QFD

To conduct QFD on an image $I(x, y)$, the Cartesian coordinate system is firstly transformed into polar coordinate system. The origin has been moved to centroid of the image. In polar coordinate system, image $I(x, y)$ is denoted as $f(r, \theta)$. Let τ be a non-negative number. For image $f(r, \theta)$, the following notations are introduced

$$\begin{aligned} u &= \int_0^{2\pi} \left(\int_0^\infty r^\tau f(r, \theta) dr \right)^4 \cos^2 \theta d\theta, & v &= \int_0^{2\pi} \left(\int_0^\infty r^\tau f(r, \theta) dr \right)^4 \sin^2 \theta d\theta, \\ w &= \int_0^{2\pi} \left(\int_0^\infty r^\tau f(r, \theta) dr \right)^4 \sin \theta \cos \theta d\theta, & \varepsilon &= \sqrt{(u - v)^2 + 4w^2}, \\ m &= \varepsilon (\sqrt{u + v + \varepsilon} - \sqrt{u + v - \varepsilon}), & p &= \frac{\sqrt{2}m(u + \sqrt{uv - w^2})}{2(u - v)^2 + 8w^2}, \\ q &= \frac{\sqrt{2}mw}{2(u - v)^2 + 8w^2}, & h &= \frac{\sqrt{2}m(v + \sqrt{uv - w^2})}{2(u - v)^2 + 8w^2}. \end{aligned}$$

Furthermore, let

$$\beta(\theta) = \frac{1}{\sqrt{(p \cos \theta + q \sin \theta)^2 + (q \cos \theta + s \sin \theta)^2}}, \quad (6)$$

$$\Gamma(\theta) = H(\theta)\pi + \arctan \frac{q \cos \theta + s \sin \theta}{p \cos \theta + q \sin \theta}, \quad (7)$$

where

$$H(\theta) = 1 - \frac{1}{2}[sign(q \cos \theta + s \sin \theta) + sign((p \cos \theta + q \sin \theta)(q \cos \theta + s \sin \theta))].$$

Here “sign” denotes the signum function.

Based on the above notations, the *central radial transform* (CRT) of image $f(r, \theta)$ is defined as follows

$$g^\tau(\theta) = \beta(\theta) \left(\int_0^\infty r^\tau f(r, \Gamma(\theta)) dr \right)^{\frac{1}{\tau+1}}. \quad (8)$$

It is noted that $g^\tau(\theta)$ is a single-valued function. The set $\{(g^\tau(\theta) \cos \theta, g^\tau(\theta) \sin \theta) | \theta \in [0, 2\pi]\}$ is a closed curve in Cartesian coordinate system. Hence, by CRT, any object is converted into a closed curve $g^\tau(\theta)$. In this paper, $g^\tau(\theta)$ is utilized as the shape signature of image $f(r, \theta)$. Discrete Fourier transform is conducted on $g^\tau(\theta)$

$$G_k = \frac{1}{N} \sum_{\theta=0}^{N-1} g^\tau(\theta) e^{-i \frac{2\theta\pi k}{N}}, \quad k = 0, 1, 2, \dots, N-1.$$

The *quasi Fourier descriptors* (QFDs) are defined as follows

$$Q_k = \frac{|G_{k+2}|}{|G_1|}, \quad k = 0, 1, 2, \dots, N-3. \quad (9)$$

3.2 Properties of QFD

It will be shown that QFDs are affine invariants. The difference of QFD with other methods is also discussed.

3.2.1 Invariance to Affine Transform

CRT defined in Eq. (8) eliminates shearing in affine transform. This property will be shown in the following theorem. This theorem can be proven similar to the proof of Theorem 1 in [19]. Here we omit the proof.

Theorem 1. *For an image $f(r, \theta)$ and its affine transformed image $\tilde{f}(\tilde{r}, \tilde{\theta}), g^\tau(\theta)$ and $\tilde{g}^\tau(\tilde{\theta})$ are their CRTs respectively. Then $g^\tau(\theta)$ can be derived by a scaling and a rotation of $\tilde{g}^\tau(\tilde{\theta})$.*

This theorem shows that CRT in Eq. (8) converts affine transform into similarity transform. $\beta(\theta)$ in Eq. (6) and $\Gamma(\theta)$ in Eq. (7) eliminate shearing in affine transform. Here, we only present the result for a binary image shown in Fig. 1 (a). Results for other binary images and gray scale images are similar. Figure 1(a) is an image of Chinese character ‘Zao’. Figure 1(d) is an affine transformed image of Fig. 1 (a). Figure 1 (b) is the closed curve $g^{0.5}(\theta)$ derived from Fig. 1 (a), and Fig. 1(e) is the closed curve $\tilde{g}^{0.5}(\tilde{\theta})$ derived from Fig. 1(d) respectively. It can be observed that Fig. 1(b) is a scaled and rotated version of Fig. 1 (e). Shearing has been eliminated. Affine transform has been converted into similarity transform.

QFDs are derived by applying FD on CRT. It follows from Theorem 1 that CRT in Eq. (8) has eliminated shearing in affine transform. On the other hand, FD is invariant to scaling and rotation. The origin has been moved to the centroid. Translation has also been eliminated. Therefore, QFDs defined in Eq. (9) are invariant to affine transform.

3.2.2 Comparison with Some Related Methods

The proposed QFD is a FD-based method. It combines contour-based and region-based techniques together. Here, we provide a comparison of QFD with some related methods.

- Comparison with traditional FD (a classical contour-based method)

Traditional FD is based on boundary derived from the object. Hence, their performance is strongly dependant on the extracted boundary. Furthermore, FD cannot get the internal structure of an object. On the contrary, the proposed QFD is based on CRT. Any object can be converted into a closed curve. For example, Chinese character ‘Zao’ in Fig. 1(a) consists of several components. It is hard for traditional FD to extract invariant features from this object. Whereas, QFD can be applied to the Chinese character ‘Zao’ in Fig. 1 (a).

- Comparison with *generic Fourier descriptor* (GFD) in [8]

In [8], Zhang et al. have proposed the notable technique: GFD. Similar to the proposed QFD, GFD is also a region-based method. But GFD can only be used to extract features invariant to similarity transform. In comparison with the proposed QFD, GFD cannot be utilized to extract features invariant to shearing.

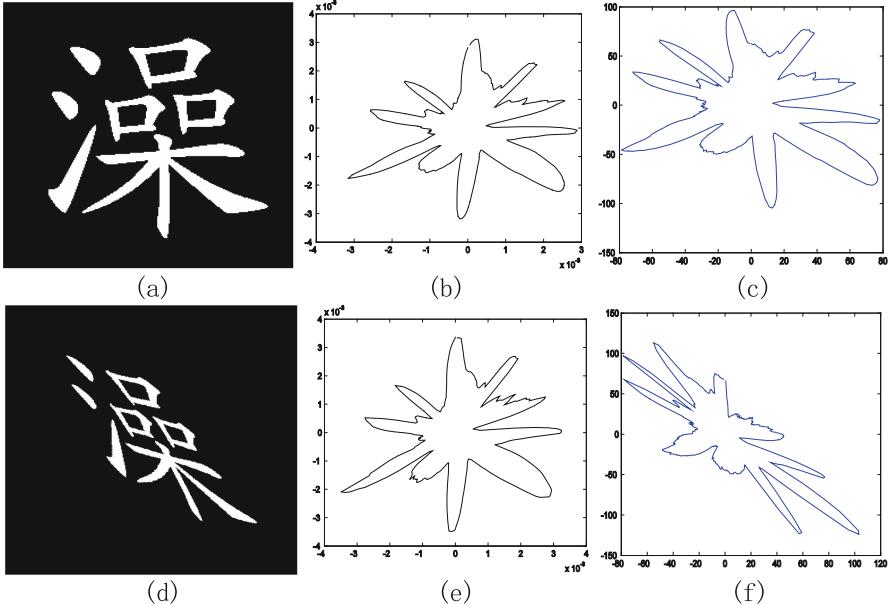


Fig. 1. Images and their CRTs, PRITs: (a) Chinese character “Zao”; (b) CRT of Fig. 1 (a);(c) PRIT of Fig. 1 (a); (d) affine transformed image of Fig. 1 (a); (e) CRT of Fig. 1 (d); (f) PRIT of Fig. 1 (d).

- Comparison with *affine Fourier descriptor*(AFD) in [3] (a contour-based method)

By a complex mathematical analysis, Arbter et al. [3] propose a set of normalize descriptors, which are called AFDs that are invariant to affine transform. In comparison with the proposed QFD, AFD is only a contour-based technique. Furthermore, parameterization is a necessary step to establish a one-to-one relation between the original contour and the transformed contour. On the contrary, the proposed QFD is a region-based technique. It can be utilized to extract features from objects with several components. Furthermore, shearing has been eliminated by CRT defined with Eq. (8). The one-to-one relation can be established without parameterization.

- Comparison with PRIT in [18]

In [18], Huang et al. propose PRIT as follows

$$p^\tau(\theta) = \left(\int_0^\infty r^\tau f(r, \theta) dr \right)^{\frac{1}{\tau+1}}. \quad (10)$$

Parameterization is conducted on the derived $p^\tau(\theta)$ to establish the one-to-one relation. Then, stationary wavelet transform is utilized to extract affine invariant features. Figure 1(c) is the PRIT of Fig. 1 (a) for $\tau = 0.5$, and Fig. 1(f) is the PRIT of Fig. 1 (d) for $\tau = 0.5$. It can be observed that Fig. 1(f) is an affine transformed version of Fig. 1(c). Shearing has not been eliminated. In comparison with PRIT, two factors $\beta(\theta)$ and $\Gamma(\theta)$ have been introduced in CRT. Shearing in affine transform has been

eliminated. Consequently, the proposed QFD needs not parameterization. The extra transform (wavelet transform, etc.) has also been avoided.

- Comparison with central projection descriptor without parameterization in [20]

In [20], Yuan et al. propose central projection descriptor which does not need parameterization for the establish of one-to-one relation. This technique can be viewed as a special case of the proposed QFD. In fact, if we set $\tau = 0$ in Eq. (8), then the proposed QFD is the same as the descriptor presented in [20]. It has been shown in [18] that the absence of radial factor may cause lose of information along radial direction in the image. The radial factor r^τ in Eq. (8) introduces information along radial direction in the image.

4 Experimental Results

In this section, the performance of QFD is tested with experiments. In Subsect. 4.1, the affine invariance of QFD is verified with several similar Chinese characters. In Subsect. 4.2, the proposed QFD is employed for pattern classification tasks. In Subsect. 4.3, the robustness of the proposed invariants to noise is tested. Furthermore, QFDs are compared with several region-based methods. For convenience, τ in Eq. (5) is set to be 0.5 in the following experiments. Results on other τ are similar.

Two databases shown in Fig. 2 are employed as the testing images. Figure 2(a) includes 20 Chinese characters, and is utilized to test performance of the proposed QFD on binary images. These Chinese characters are with *regular script front*, and each of these characters has a size of 256×256 . Figure 2(b) includes gray scale images in Coil-20 database [21]. Each of these images is of 128×128 .

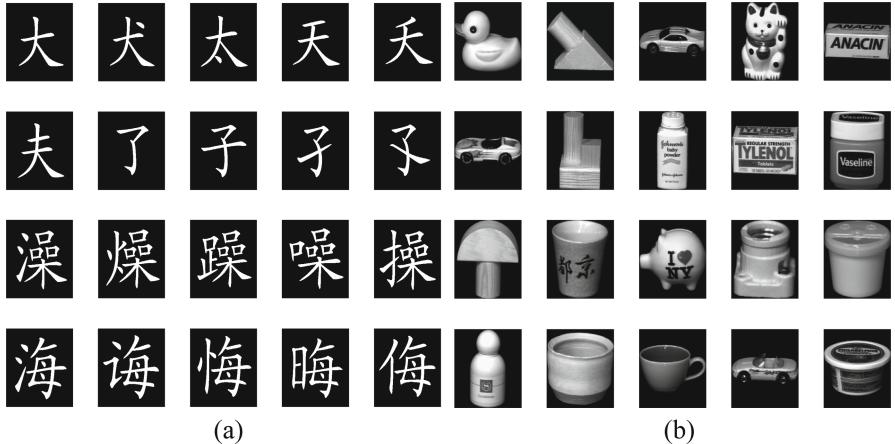


Fig. 2. Testing images: (a) 20 Chinese characters; (b) 20 Gy scale images in [21].

Suppose that features extracted from image f and g are respectively denoted as $\{V_k^f\}$ and $\{V_k^g\}$. Then similarity between them is defined as the following correlation

coefficient:

$$\text{Sim}(V_k^f, V_k^g) = \frac{\sum_k V_k^f V_k^g}{\sqrt{\sum_k (V_k^f)^2} \sqrt{\sum_k (V_k^g)^2}}. \quad (11)$$

The proposed QFD is compared with AMIs and MSA. For AMIs, the first three invariants in [14] are used. For MSA, 29 invariants as reported in [16] are used.

4.1 Testing of Invariance

In this subsection, the affine invariance of QFD is validated. The first six Chinese characters in Fig. 2(a) are employed as the testing images. Each of these images is transformed one time. Results on other transforms and other images are similar. N in Eq. (9) is set to be 360. Table 1 lists correlation coefficients for the used six Chinese characters and their transformed images with the proposed QFD. From Table 1, it can be observed that correlation coefficients between the original image and its affine transformed images are larger than those between the affine transformed image and other images. It may even be difficult for a human being to recognize some of these six characters under affine transformation. By QFD, these characters can be correctly recognized. Therefore, the proposed QFD is invariant to affine transform.

Correlation coefficients using AMIs and MSA have also been calculated. Results are listed in Table 2 and Table 3 respectively. It can be observed from Table 2 that these Chinese characters can hardly be recognized by AMIs. From Table 3, it can be seen that some characters cannot be distinguished with MSA.

Table 1. Correlation coefficients of six similar Chinese characters in Fig. 2 (a) and their affine transformed images with the proposed QFD

	大	卡	太	夭	火	夫
大	0.9938	0.9361	0.9104	0.9254	0.8959	0.9153
犬	0.9237	0.9948	0.9325	0.9585	0.9475	0.9233
太	0.8845	0.9098	0.9943	0.9093	0.9171	0.9085
夭	0.9141	0.9156	0.9163	0.9891	0.9457	0.8970
火	0.8866	0.9334	0.9151	0.9681	0.9926	0.9089
夫	0.9287	0.9153	0.9396	0.9254	0.9178	0.9870

Table 2. Correlation coefficients of six similar Chinese characters in Fig. 2 (a) and their affine transformed images with AMIs

	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3. Correlation coefficients of six similar Chinese characters in Fig. 2 (a) and their affine transformed images with MSA

	1.0000	0.9998	1.0000	0.9986	0.9992	0.9974
	0.9998	1.0000	0.9998	0.9993	0.9996	0.9983
	1.0000	0.9998	1.0000	0.9986	0.9993	0.9973
	0.9985	0.9992	0.9985	1.0000	0.9996	0.9996
	0.9992	0.9996	0.9992	0.9997	1.0000	0.9990
	0.9973	0.9981	0.9972	0.9996	0.9989	1.0000

4.2 Pattern Classification

The proposed QFD is further utilized for pattern classification tasks. Images in Fig. 2 are employed for the testing. Each of these images is transformed by the following

transformation matrix as in [2]:

$$T = l \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} a & b \\ 0 & 1/a \end{pmatrix}. \quad (12)$$

Here, we choose $l \in \{0.8, 1.2\}$, $\theta \in \{0^\circ, 72^\circ, 144^\circ, 216^\circ, 288^\circ\}$, $b \in \{-1.5, -1, -0.5, 0, 0.5, 1, 1.5\}$ and $a \in \{1, 2\}$. Consequently, each image in the testing databases is transformed 140 times. As a result, 2800 tests run for every database (20×140).

The test image is classified as the most similarity pattern. The larger correlation coefficient between two objects, these two objects have higher similarity. The classification accuracy is defined as follows

$$\eta = \frac{N_c}{N_t} \times 100\%,$$

where N_c denotes the number of correctly classified images, N_t denotes a total number images applied in the test.

Experiments on 20 Chinese characters in Fig. 2(a) and their affine transformations show that the proposed QFD can achieve 99.46% accurate classification. The classification accuracies for AMIs and MSA are 96.64% and 99.25% respectively. Experiments have also been conducted on gray images in Fig. 2(b) and their affine transformations. The classification accuracy for the proposed QFD comes to 99.96%, Accuracies for AMIs and MSA are 100.00% and 95.18% respectively.

4.3 Robustness to Noise

In this subsection, the robustness of QFD to noise is tested by pattern classification task.

20 Chinese characters in Fig. 2(a) are used to test the performance of QFD on binary images. All methods are tested by adding Salt & Pepper noise with intensities varying from 0 to 0.025. Results are presented in Fig. 3. It can be observed from Fig. 3 that all these methods have a high discrimination in noise-free condition. The proposed QFD performs a little better than AMIs and MSA. In noisy condition, classification accuracy of the proposed method and MSA decrease much more slowly than AMIs. MSA performs a little better than the proposed QFD at noise intensity 0.005, but its accuracies are less for high noise levels.

20 gray scale images in [21] are utilized to test the performance of QFD on gray scale images. All these methods are tested by adding Gaussian noise with intensities varying from 0 to 0.005. Figure 4(a) shows the classification accuracies for different noise intensities. It can be observed from Fig. 4(a) that AMIs and MSA have very low recognition accuracies. In high noise intensity, the proposed QFD still achieves satisfied classification accuracy. In addition, all these methods are also tested by adding Salt & Pepper noise with intensities varying from 0 to 0.025. Figure 4(b) shows the results. It can also be observed from Fig. 4(b) that the proposed QFD performs better than AMIs and MSA.

These results indicate that the proposed method is much more robust to noise than AMIs and MSA. The reason lies in that the integral along radial direction in CRT (see

Eq. (8)) eliminates the effect of additive noise. Similar to the discussion in [18] for PRIT in Eq. (10), it can be shown that CRT is much robust to noise.

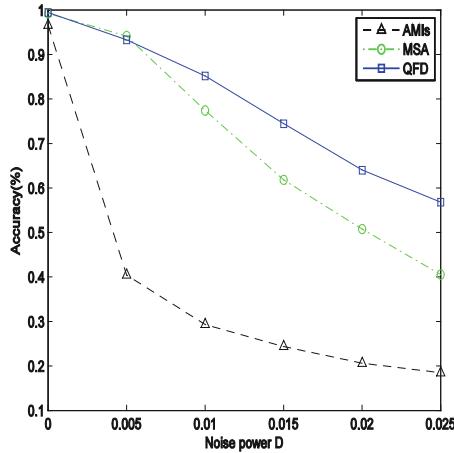


Fig. 3. Performance on binary images in Fig. 2(a) for Salt & Pepper noise.

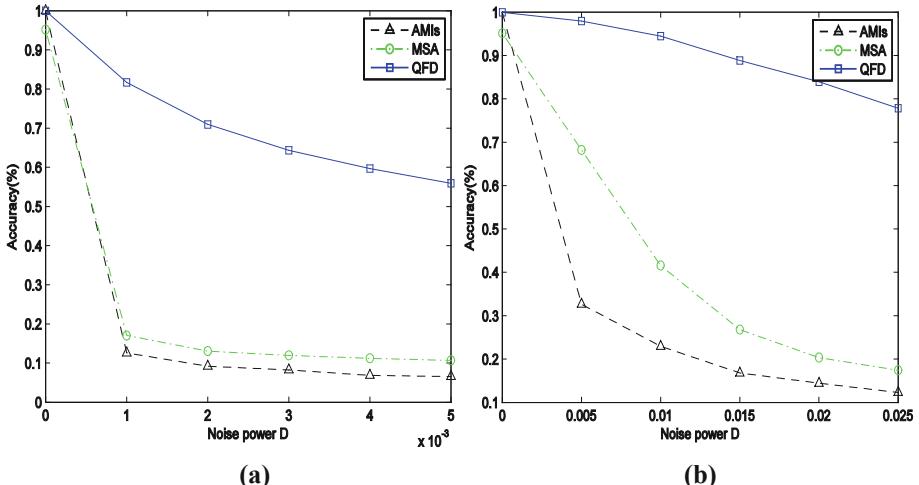


Fig. 4. Performance on gray scale images in Fig. 2(b): (a) for Gaussian noise; (b) for Salt & Pepper noise.

5 Conclusions

Great scientific attentions have been attracted for the extraction of affine invariant features. Contour-based methods and region-based methods both have their advantages and

disadvantages. To keep the data reduction of contour-based technique, and to keep the high accuracy of region-based methods, CRT is proposed to combine them together. By CRT, any object is converted into a closed curve. Shearing is eliminated. Consequently, QFDs are derived by conducting FD on CRT. In other words, the proposed QFD is equivalent to applying FD on CRT. As a result, the obtained QFDs are invariant to affine transform.

The proposed QFD is strongly dependant on accuracy for the calculated centroid. Large occlusion causes a bad estimation of centroid. As a result, QFDs are much imprecise for occlusion. In addition, the choice of τ in CRT (Eq. (8)) is a problem that should be addressed.

Acknowledgements. This work was supported by the National Natural Science Foundation of China, (grant nos.: 51705120). This work was also supported by the 24st batch of university students' extracurricular academic research fund of Soochow University.

References

- Flusser, J., Suk, T., Zitová, B.: 2D and 3D Image Analysis by Moments. Wiley, Hoboken (2016)
- Khalil, M.I., Bayoumi, M.M.: A dyadic wavelet affine invariant function for 2D shape recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(10), 1152–1163 (2001)
- Arbter, K., Snyder, W.E., Burkhardt, H., Hirzinger, G.: Application of affine-invariant Fourier descriptors to recognition of 3-D objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(7), 640–647 (1990)
- Li, H., Jin, X., Yang, N., Yang, Z.: The recognition of landed aircrafts based on PCNN model and affine moment invariants. *Pattern Recogn. Lett.* **51**, 23–29 (2014)
- Gishkori, S., Mulgrew, B.: Pseudo-Zernike moments based sparse representations for SAR image classification. *IEEE Trans. Aerospace Electron. Syst.* **55**(2), 1037–1044 (2019)
- Hao, Y., Li, Q., Mo, H., Zhang, H., Li, H.: AMI-Net: convolution neural networks with affine moment invariants. *IEEE Signal Process. Lett.* **25**(7), 1064–1068 (2018)
- Gong, M., Hao, Y., Mo, H., Li, H.: Naturally combined shape-color moment invariants under affine transformations. *Comput. Vis. Image Underst.* **162**, 46–56 (2017)
- Zhang, D.S., Lu, G.J.: Review of shape representation and description techniques. *Pattern Recogn.* **37**, 1–19 (2004)
- Zahn, C.T., Roskies, R.Z.: Fourier descriptors for plane closed curves. *IEEE Trans. Comput. C* **21**(3), 269–281 (1972)
- Zhang, D.J., Lu, G.S.: Shape-based image retrieval using generic Fourier descriptor. *Signal Process. Image Commun.* **17**(10), 825–848 (2002)
- Yang, C., Yu, Q.: Multiscale Fourier descriptor based on triangular features for shape retrieval. *Signal Process. Image Commun.* **71**, 110–119 (2019)
- El-ghazal, A., Basir, O., Belkasim, S.: Invariant curvature-based Fourier shape descriptors. *J. Vis. Commun. Image Represent.* **23**(4), 622–633 (2012)
- Yang, C., Wei, H., Yu, Q.: A novel method for 2D non rigid partial shape matching. *Neurocomputing* **275**, 1160–1176 (2018)
- Flusser, J., Suk, T.: Pattern recognition by affine moment invariants. *Pattern Recogn.* **26**, 167–174 (1993)
- Yang, Z., Cohen, F.: Cross-weighted moments and affine invariants for image registration and matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **21**(8), 804–814 (1999)

16. Rahtu, E., Salo, M., Heikkila, J.: Affine invariant pattern recognition using multiscale autoconvolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(6), 908–918 (2005)
17. Lan, R.S., Yang, J.W., Jiang, Y., Colin, F., Song, Z.: Whitening central projection descriptor for affine-invariant shape description. *IET Image Proc.* **7**(1), 81–91 (2013)
18. Huang, Y.D., Yang, J.W., Li, S.S., Du, W.Z.: Polar radius integral transform for affine invariant feature extraction. *Int. J. Wavelets Multiresolut. Inf. Process.* **15**(01), 640–647 (2017)
19. Yang, J.W., Lu, Z.D., Tang, Y.Y., Yuan, Z., Chen, Y.J.: Quasi Fourier-Mellin transform for affine invariant features. *IEEE Trans. Image Process.* **29**, 4114–4129 (2020)
20. Yuan, Z., Yang, J.W.: Central projection descriptor without parameterization. *J. Optoelectron. Laser*, **30**(4), 434–441 (2019). (in Chinese)
21. Nene, S.A., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-20), Tech. Report. CUCS-005-96, February 1996. The database can be downloaded from: <http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>.



A New PM2.5 Concentration Predication Study Based on CNN-LSTM Parallel Integration

Chaoxue Wang, Zhenbang Wang^(✉), Fan Zhang, and Yuhang Pan

School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
ae@wangzhenbang.com

Abstract. Prevention and control of haze is the hot topic in the study of air quality, and PM2.5 concentration prediction is one of the keys in the haze prevention and control. This paper proposes a new method of integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) in parallel to predict PM2.5 concentration. This method can learn the spatial and temporal characteristics of data at the same time, and has powerful integrated learning capabilities. Taking the prediction of PM2.5 concentration in Xi'an area as an example, the method in this article is compared with the method in relevant authoritative literature. The experimental results show that the method in this article has better prediction effect and is a more competitive deep learning prediction model.

Keywords: PM2.5 concentration predication · Convolutional Neural Networks (CNN) · Long Short-Term Memory Networks (LSTM) · Parallel integrated learning

1 Introduction

As the main component of smog, PM2.5 is an important source of air pollution. Due to its small size, it can adhere to the deep respiratory tract and affect blood circulation by penetrating lung cells, posing a great threat to people's health [1]. Studies have shown that the increase in PM2.5 concentration increases the risk of people suffering from respiratory and cardiovascular diseases, and attenuates people's lung function [2, 3]. Therefore, accurate prediction of PM2.5 concentration, timely understanding of air pollution levels and preventive measures are of great significance to human health and environmental protection.

The current PM2.5 concentration prediction is based on artificial intelligence methods, among which deep learning is the current hot method [4, 5]. Kuremoto et al. [6] used a deep network composed of two Restricted Boltzmann Machines (RBMs) for timing prediction, and used CATS (Competition on Artificial Time Series) benchmarks and original data to prove that RBMs are better than the regression moving average model (Auto Regressive Integrated Moving Average, ARIMA). Ong et al. [7] used Deep Recurrent Neural Network (DRNN) to predict the concentration of air pollutants, and it is more effective than RBMs. Feng et al. [8] combined Random Forest (RF) and RNN to

predict the PM2.5 concentration of Hangzhou in the next 24 h. When the time interval of the input data increases, the RNN network may have gradient disappearance or gradient explosion, and the long-term predictive ability is limited. The long short-term memory network (Long Short-Term Memory Networks, LSTM) can effectively solve the problem of the RNN network. Question [9]. Liu et al. [10] constructed a PM2.5 concentration prediction model based on LSTM recurrent neural network, which can effectively solve the long-term dependence of air pollutants, and the effect is better than RNN. The input of LSTM is historical data, and there is no future data. The bi-directional LSTM (Bi-Directional LSTM, Bi-LSTM) can analyze the time series of historical data and future data, and the ability to learn long-term dependence on information is better [11]. Zhang et al. [12] used Bi-LSTM to construct a PM2.5 concentration prediction model based on a self-encoding network, which is more accurate for long-term prediction of PM2.5 concentration. Convolutional Neural Networks (CNN) are powerful in spatial data processing [13]. Through satellite image analysis, this type of method has also been used to predict the concentration of pollutants. Huang Weizheng et al. [14] used CNN to extract features from remote sensing images to predict PM2.5 concentration. Due to the spatiotemporal nature of air pollutants, Huang et al. [15] used the CNN-LSTM tandem combination method to predict the future PM2.5 concentration hour by hour. Li et al. [16] used the method of combining CNN and LSTM in series with the attention mechanism to predict the hourly PM2.5 concentration in the future. Zhang and Zhao et al. [17] constructed a series PM2.5 concentration prediction model based on CNN-LSTM, added multi-site pollutants and meteorological data, first used CNN to extract features of PM2.5 concentration factors, and then extracted the features Importing into LSTM to predict PM2.5 concentration effectively improves the accuracy of PM2.5 concentration prediction.

On the basis of predecessors, especially inspired by the literature [15–17], this paper proposes a method for PM2.5 concentration prediction by integrating CNN and LSTM in parallel. This method can simultaneously learn the spatial and temporal characteristics of data, and has powerful predictive modeling capabilities. This article takes the PM2.5 concentration prediction in Xi'an as an example, uses Correlation Coefficient (Corr) and Root Mean Square Error (RMSE) as evaluation indicators, and compares the methods in this article with those in relevant authoritative documents. Compared with simulation experiments, the experimental results prove the effectiveness and advancement of the method in this paper.

2 A New Parallel Integrated Prediction Method of CNN and LSTM

CNN can effectively extract the spatial features of the data, and LSTM can analyze the time series of the data. It is a better method to integrate CNN and LSTM in series to predict PM2.5 concentration. Based on the literature [15–17], this paper first designed a new CNN model and LSTM model to learn the spatial and temporal characteristics of PM2.5, and obtained two sets of PM2.5 concentration prediction values, and then the two sets of predicted values are input to the fully connected network model for fusion, and finally the final predicted value that can reflect both the spatial and temporal characteristics of PM2.5 concentration is obtained.

2.1 Parallel Integrated Prediction Model Framework

Figure 1 shows the CNN and LSTM parallel integrated prediction model framework proposed in this paper. First, obtain air pollutant data and meteorological data, including PM2.5, PM10, SO2, NO2, CO, O3, temperature, humidity, air pressure, dew point, wind speed, The wind direction is input into the CNN model and the LSTM model respectively, and then the two sets of PM2.5 concentration prediction values obtained by the CNN model and the LSTM model are combined and input into the fully connected network model for further learning to realize the prediction of PM2.5 concentration.

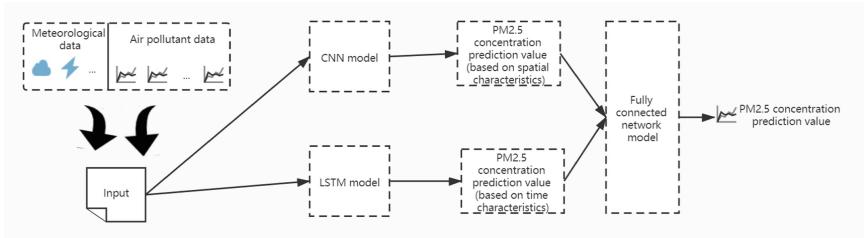
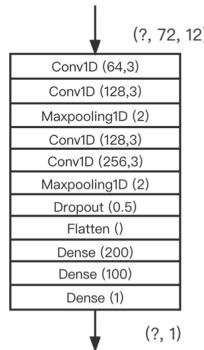


Fig. 1. CNN-LSTM parallel integration predication model framework

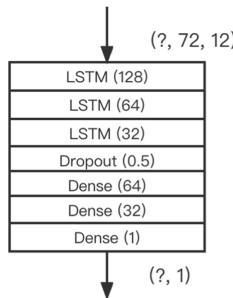
2.2 CNN Model

The specific design of the CNN model in Fig. 1 is shown in Fig. 2. The input of the model is the air pollutant data and meteorological data in the past 72 h, including PM2.5, PM10, SO2, CO, NO2, O3, temperature, dew point, humidity, Air pressure, wind speed, wind direction, the model has 11 layers. The first layer of the CNN model is the convolutional layer Conv1D (64,3), its convolution kernel is 64, the convolution kernel size is 3; the second layer is the convolutional layer Conv1D (128,3), its convolution core There are 128 product cores and the convolution core size is 3; the third layer is the maximum pooling layer MaxPooling1D (2), the size is 2; the fourth layer is the convolution layer Conv1D (128,3), and its convolution core is 128, the convolution kernel size is 3; the fifth layer is the convolution layer Conv1D (256,3), its convolution kernel is 256, the convolution kernel size is 3; the sixth layer is the maximum pooling layer MaxPooling1D (2) The size is 2. The first six layers mainly extract the spatial features of the input data through the convolutional layer, and the maximum pooling layer compresses the extracted spatial features. The seventh layer is the Dropout (0.5) layer, and the probability of neuron discarding is 0.5. This layer is used to eliminate the over-fitting phenomenon of the model; the eighth layer Flatten () expands the output of the seventh layer in one dimension with default parameters. The ninth layer Dense (200) is a fully connected layer composed of 200 neurons, which integrates and learns the obtained spatial features; the tenth layer Dense (100) is a fully connected layer composed of 100 neurons, which further improves the model Non-linear expression ability; the eleventh layer Dense (1) is a fully connected layer composed of 1 neuron to output the predicted value of PM2.5 concentration.

**Fig. 2.** CNN model

2.3 LSTM Model

The specific design of the LSTM model in Fig. 1 is shown in Fig. 3. The input of the model is the air pollutant data and meteorological data in the past 72 h, including PM2.5, PM10, SO2, CO, NO2, O3, temperature, dew point, Humidity, air pressure, wind speed, wind direction, the model has 7 layers. The first three layers are all LSTM layers, with the number of neurons being 128, 64, and 32 respectively, and its function is mainly to extract the time characteristics of the input data. The fourth layer is the Dropout layer, and the discard rate of neurons is 0.5, which is used to eliminate the overfitting of the model. The last three layers are all fully connected layers with parameters of 64, 32, and 1, respectively, which are used to obtain the predicted value of PM2.5 concentration.

**Fig. 3.** LSTM model

2.4 LSTM Model

The fully connected network model is shown in Fig. 4 and has 4 layers. The input data is a combination of PM2.5 concentration prediction values obtained by the CNN model and the LSTM model. The data format is shown in Fig. 5. The first layer Dense (50) is a fully connected layer composed of 50 neurons. This layer performs integrated

learning on the combined data obtained. The second layer is the Dropout layer. The probability of neuron discarding is 0.2, which is used to eliminate the model's overload fitting phenomenon, the third layer Dense (30) is a fully connected layer composed of 30 neurons, which further improves the nonlinear expression ability of the model, and the fourth layer Dense (1) is a fully connected layer composed of 1 neuron to output the predicted value of PM2.5 concentration.

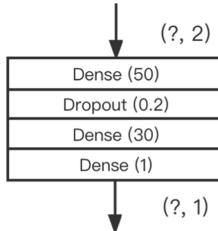


Fig. 4. Fully connected network model

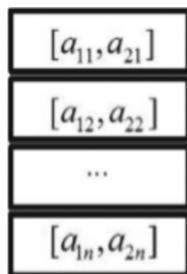


Fig. 5. Data set format of the model

3 Experiments and Results

The programming language is Python3.5, the development platform is PyCharm 2020, the deep learning framework is an open-source framework Keras based on Tensorflow, and the operating system is Windows 10.

3.1 Experimental Data

Obtained hour by hour PM2.5, PM10, SO₂, NO₂, CO, O₃ data of 6 air pollutants in Xi'an from 0:00 on January 1, 2015 to 23:00 on May 31, 2018 through the China Environmental Monitoring Website And 6 meteorological data of temperature, humidity, pressure, dew point, wind speed, and wind direction are used as experimental data. 60% of the samples in the data set are the training set, 20% are the validation set, and 20% are the test set.

Data Preprocessing

The missing data in the data set is complemented by the mean value method; because the dimensions of each decision factor are inconsistent, the value range is very different, so the data is standardized before the experiment.

Time Step Processing

The time series of the input data $X = (X_{t-72}, \dots, X_{t-1}, X_t)$ is set to be a 12-dimensional vector representing the input data at time t, including six air pollutants data of PM2.5, PM10, SO2, NO2, CO, and O3, as well as temperature, dew point, humidity, six meteorological data of air pressure, wind speed, and wind direction; the length of the time series is 72, and the unit is hour. The time series for outputting the predicted target value $Y = (Y_{t+1}, Y_{t+2}, \dots, Y_{t+24})$ is assumed to have a length of 24, Y_{t+1} represents the predicted value of the PM2.5 concentration of the model at time t + 1.

3.2 Experimental Results

In the course of the experiment, four models including CNN alone, LSTM alone, CNN-LSTM in series, and CNN-LSTM in parallel were used to predict PM2.5 concentration. The evaluation index adopts the root mean square error RMSE and the correlation coefficient Corr, the expressions of which are shown in formula 1 and formula 2.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

$$Corr = \frac{Cov(y_i, \hat{y}_i)}{\sqrt{Var[y_i] * Var[\hat{y}_i]}} \quad (2)$$

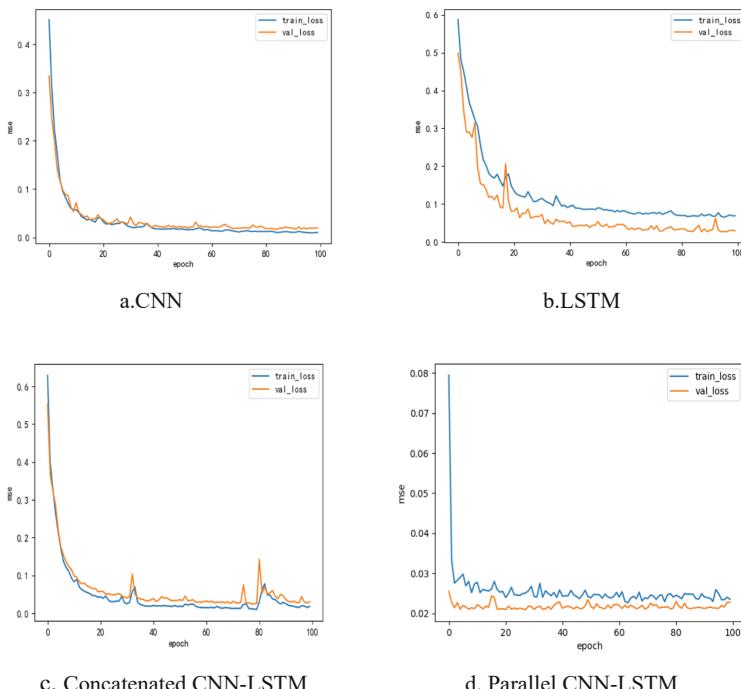
y_i represents the true value of PM2.5 concentration, \hat{y}_i represents the predicted value of PM2.5 concentration in the model, and $Var[y_i]$ represents the variance of the true value and $Var[\hat{y}_i]$ predicted value of PM2.5 respectively, and $Cov(y_i, \hat{y}_i)$ represents the covariance of the true value and predicted value.

The comparison between CNN, LSTM, series CNN-LSTM and the CNN, LSTM, series CNN-LSTM and parallel CNN-LSTM designed in this paper in [17] is shown in Table 1. It can be seen from Table 1 that not only the three models designed in this paper have a smaller RMSE and a slightly better fit than the corresponding model prediction results in the literature [17], and the parallel CNN-LSTM is the comprehensive performance of the prediction results of all models. the best.

Table 1. RMSE and Corr of each model

Model	RMSE (epochs = 100)	Corr (epochs = 100)
CNN [17]	30.66	0.980
LSTM [17]	17.95	0.950
Concatenated CNN-LSTM [17]	14.3	0.970
This text CNN	8.57	0.990
This text LSTM	12.77	0.977
This article is concatenated with CNN-LSTM	8.89	0.989
This article parallels CNN-LSTM	8.24	0.991

The loss function of the four models of CNN, LSTM, series CNN-LSTM, and parallel CNN-LSTM designed in this paper during the training process is Mean Square Error (MSE), which increases with the number of iterations on the training set and validation set. The changes are shown in Fig. 6a, b, c, and d. It can be seen from Fig. 6 that the parallel CNN-LSTM model not only has the smallest loss function, but also has the fastest convergence speed, indicating that the performance of the model is relatively optimal.

**Fig. 6.** Loss function iteration diagram for the model

The fitting curves of the true and predicted values of the four models designed in this paper, CNN, LSTM, series CNN-LSTM, and parallel CNN-LSTM on the same test data set, are shown in Fig. 7a, b, c, and d, respectively. It can be found in Fig. 7 that the parallel CNN-LSTM model has the best fitting effect.

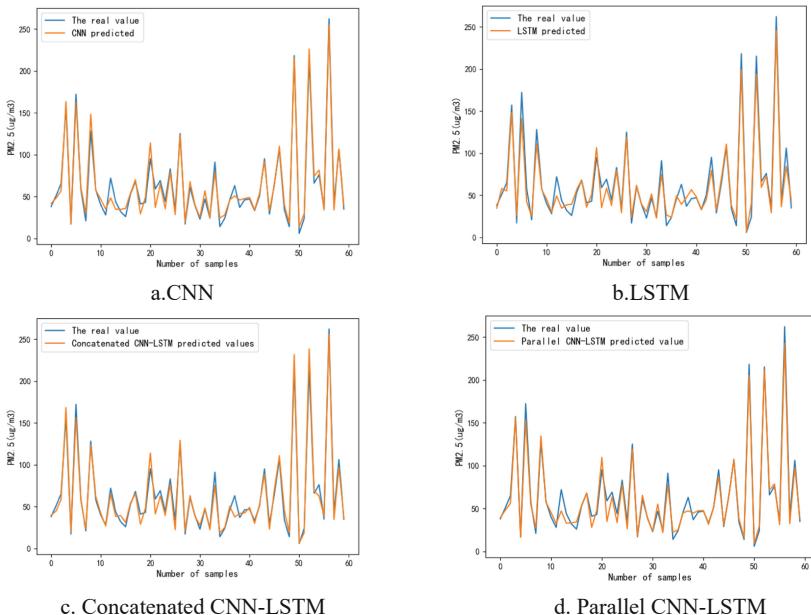


Fig. 7. Prediction fit curve of the model

4 Summary

This paper proposes a new parallel integrated model based on CNN and LSTM to predict the concentration of PM2.5. This model can simultaneously learn the spatial and temporal characteristics of the data, and has a powerful integrated learning ability. Through the simulation experiment of PM2.5 concentration prediction in Xi'an area, it shows that the model proposed in this paper has better performance than the model in the related literature compared in the experiment.

In future work, for the prediction of PM2.5 concentration, more influencing factors can be considered, such as traffic impact, industrial governance, economic conditions, etc., and data sets from multiple sites can be added to consider the difference between different sites. The interaction between air pollutants.

References

1. Kampa, M., Castanas, E.: Human health effects of air pollution. Environ. Pollution **151**(2), 362–367 (2008)

2. Kappos, A.D., Bruckmann, P., Eikmann, T., et al.: Health effects of particles in ambient air. *Int. J. Hyg. Environ. Health* **207**(4), 399–407 (2004)
3. Lim, S.S., Vos, T., Flaxman, A.D., et al.: A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factors clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *Lancet* **380**(9859), 2224–2260 (2012)
4. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **2015**(61), 85–117 (2015)
5. Deng, L., Yu, D.: Deep learning: method and applications. *Found. Trends Signal Process.* **7**(3–4), 197–387 (2013)
6. Kuremoto, T., Kimura, S., Kobayashi, K., et al.: Time series forecasting using a deep belief network with restricted Boltzmann machines. *Neurocomputing* **137**(2014), 47–56 (2014)
7. Ong, B.T., Sugiura, K., Zettsu, K.: Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. *Neural Comput. Appl.* **27**(2016), 1553–1566 (2016)
8. Feng, R., Zheng, H.J., Gao, H., et al.: Recurrent neural network and random forest for analysis and accurate forecast of atmospheric pollutants: a case study in Hangzhou, China. *J. Clean. Prod.* **2019**(231), 1005–1015 (2019)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
10. Liu, X.D., Liu, Q., Zou, Y.Y., et al.: A self-organizing LSTM-Based approach to PM2.5 forecast. In: Proceedings of International Conference on Cloud Computing & Security, pp. 683–693 (2018)
11. Lin, B.Y., Xu, F.F., Luo, Z.Y., et al.: Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media. In: Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 160–165 (2017)
12. Zhang, B., Zhang, H.W., Zhao, G.M., et al.: Constructing a PM2.5 concentration predication model by combining auto-encoder with Bi-LSTM neural networks. *Environ. Modell. Softw.* **124**(2020), 104600–104610 (2020)
13. LeCun, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to recognition. In: Proceedings of the IEEE, pp. 2278–2324 (1998)
14. Huang, W.: Research on Spatiotemporal Evolution Predication Model of Haze Based on Convolution Neural Network. University of Electronic Science and Technology of China, Chengdu (2018). (in Chinese)
15. Huang, C.J., Kuo, P.H.: A deep CNN-LSTM model for predication matter (PM2.5) forecasting in smart cities. *Sensors* **18**(7), 2220–2241 (2018)
16. Li, S.Z., Xie, G., Ren, J.C., et al.: Urban PM2.5 concentration predication via attention-based CNN-LSTM. *Appl. Sci.* **10**(6), 1953–1970 (2020)
17. Qin, D.M., Zou, G.J., et al.: A novel combined predication scheme based on CNN and LSTM for urban PM2.5 concentration. *IEEE Access* **7**(2019), 20050–20059 (2019)

Attached Chinese References

18. 黃伟政. 基于卷积神经网络的雾霾时空演化预测方法研究[D]. 成都: 电子科技大学, 2018



Deep Discriminant Non-negative Matrix Factorization Method for Image Clustering

Kexin Xie¹, Wen-Sheng Chen^{1,2(✉)}, and Binbin Pan^{1,2}

¹ College of Mathematics and Statistics, Shenzhen University, Shenzhen, China
`{chenws, pbb}@szu.edu.cn`

² Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060,
People's Republic of China

Abstract. Non-negative matrix factorization (NMF) is a single-layer decomposition algorithm for image data analysis and cannot reveal the intrinsic hierarchical-structure information hidden under the data. Moreover, it extracts features with weak discriminant power due to merely using unlabeled data. These flaws cause the undesirable performance of NMF in image clustering. To address the problems of NMF, this paper presents a novel deep discriminant NMF (DDNMF) approach via neural networks (NN). Our DDNMF model is established by employing the loss function of neural networks and the supervised regularization terms. The optimization problem of DDNMF is solved utilizing gradient descent method (GDM). The proposed DDNMF algorithm is proven to be convergent and stable. Some related NMF-based algorithms are selected for comparison on two facial image datasets. The experimental results confirm the effectiveness and superior performance of our DDNMF method in image clustering tasks.

Keywords: Non-negative matrix factorization · Neural networks · Deep feature extraction · Image clustering

1 Introduction

Non-negative matrix factorization [1, 2], proposed by Lee *et al.*, is a powerful tool for non-negative data processing and dimensionality reduction. NMF and its variations have variety applications in image data analysis, such as face recognition [3–5], image clustering [6–8], signal unmixing [10, 11], and so on. The goal of NMF is to find two low-rank factor matrices W and H such that their product approximates a data matrix X , namely $X \approx WH$, where three matrices X , W and H are non-negative and called data matrix, basis matrix and feature matrix, respectively. However, NMF is a single-layer feature extraction approach and cannot uncover the latent hierarchical feature of the data, while the deep feature is more useful for complex image data representation. Furthermore, NMF does not make use of the labeled data for training and is an unsupervised machine learning algorithm. It is known that label information can potentially boost the algorithm performance in classification and clustering tasks. To disclose the cascaded structure of

the data, Ahn *et al.* [9] extended NMF to a multi-layer model and developed a multiple non-negative matrix factorization (MNMF) algorithm. It repeatedly decomposes the feature matrix using two intermediate matrices and has the ability to learn a hierarchical-parts representation. Cichocki *et al.* [10] proposed a multi-layer non-negative matrix factorization algorithm using the projected gradient technique (PMNMF) for blind source unmixing. The PMNMF algorithm factorizes the feature matrix layer-by-layer via projected NMF to create a multi-layer structure. Zhao *et al.* recursively conducted the regularized NMF on the basis matrix to acquire a regularized deep non-negative basis matrix factorization (RDNBMF) algorithm for face recognition. Recently, Zeng *et al.* [8] came up with a BP neural network-based deep NMF (BPDNMF) algorithm to learn the hierarchical feature for image clustering. Nevertheless, for the MNMF, PMNMF and RDNBMF methods, the cumulative error caused by the layer-by-layer decomposition raises their overall decomposition error. This indicates that their deep decomposition is inaccurate. Also, similar to NMF, none of these methods utilizes the label information of the samples, and they are all unsupervised learning methods. In particular, neither MNMF nor PMNMF gives the proof of their convergence, which means that the stability of the algorithms cannot be guaranteed. As for BPDNMF, although it is a supervised learning method, the main drawback is that its input feature is fixed and cannot be automatically modulated by the neural network. The shortcomings mentioned above will negatively affect the performance of these NMF-based algorithms in image clustering.

To resolve the aforementioned problems encountered in NMF and deep NMF methods, this paper attempts to develop an algorithm that is not only able to build the deep hierarchical structure through the automatic coding of a neural network but also fully makes use of the class label knowledge to enhance the clustering accuracy of the deep feature. To this end, we establish a deep discriminant NMF (DDNMF) model by combining the loss of a neural network with two scatter quantities, namely intra-class scatter and total-class scatter. Minimizing the intra-class scatter while maximizing the total-class scatter will assist in improving the performance of the proposed DDNMF algorithm. The optimization problem of our DDNMF model is tackled using the gradient descent method. The DDNMF algorithm is proved to be convergent and stable. The proposed DDNMF method is evaluated on facial images for clustering. Compared with the state-of-the-art NMF-based algorithms, the results show the effectiveness and superior performance of our DDNMF approach.

The rest of this paper is organized as follows: Sect. 2 briefly introduces the BP neural network. Section 3 creates our DDNMF model and derives the corresponding algorithm. In Sect. 4, the convergence of the proposed algorithm is theoretically analyzed. Section 5 reports the experimental results. The conclusion of this paper is drawn in Sect. 6.

2 Framework of the Back Propagation Neural Network

DDNMF aims to produce a deep NMF directly from a back propagation neural network (BPNN) [12]. So, this section will briefly introduce the framework of BPNN. Details are as follows.

2.1 Forward Propagation

For a neural network, we assume L is the total number of the layers, m_i is the number of neurons at layer i ($i = 1 \dots L$), W_i and b_i are the given weight matrix and the bias at layer i , respectively, and $f(\cdot)$ is an activation function. The procedure of forward propagation is shown below:

$$\begin{aligned} a_0 &= x, \\ z_i &= W_i a_{i-1} + b_i, \\ a_i &= f(z_i), \end{aligned}$$

where x is the input data, z_i and a_i are the input and the output at layer i , respectively. If y is the output target, the neural network needs to reduce the loss function $L_{BPNN}(a_L, y)$ defined by (1).

$$L_{BPNN}(a_L, y) = \frac{1}{2} \|a_L - y\|_F^2. \quad (1)$$

2.2 Backward Propagation

After forward propagation, we can obtain z_i and a_i at each layer. Therefore, the process of backward propagation is given as follows:

$$\delta_L = f'(z_L) \odot (a_L - y),$$

for $i = L - 1, \dots, 2, 1$,

$$\delta_i = f'(z_i) \odot (W_{i+1}^T \delta_{i+1}),$$

where $\delta_i = \nabla_{z_i} L_{BPNN}(a_L, y)$.

Since $\nabla_{W_i} L(a_L, y) = \delta_i a_{i-1}^T$ and $\nabla_{b_i} L(a_L, y) = \delta_i$, we can update the weight W_i and bias b_i using the following gradient descent method.

$$W_i \leftarrow W_i - r_W \nabla_{W_i} L_{BPNN}(a_L, y), \quad b_i \leftarrow b_i - r_b \nabla_{b_i} L_{BPNN}(a_L, y),$$

where r_W and r_b are the learning rates of W and b , respectively.

3 Proposed DDNMF Approach

This section will employ BPNN and data-dispersion information to establish our DDNMF model, which is then solved using the gradient descent method to obtain the proposed DDNMF algorithm.

Let c be the number of class, n_i be the data number of class i and $n = \sum_{i=1}^c n_i$ be the total number of training data. The training data matrix X is denoted as $X = [X_1, X_2, \dots, X_c]$, where X_i is the data matrix from class i with $X_i = [x_1^{(i)}, x_2^{(i)}, \dots, x_{n_i}^{(i)}] \in R^{m \times n_i}$. In our BPNN, we randomly generate a non-negative input matrix as $H = [H_1, H_2, \dots, H_c]$ with $H = [h_1^{(i)}, h_2^{(i)}, \dots, h_{n_i}^{(i)}] \in R^{m_1 \times n_i}$, select X as the output target. Especially, we set the bias to 0 and choose activation function as $f(x) = p^{\frac{1}{L}}x$ with parameter $p > 0$. The flowchart of our BPNN is demonstrated in Fig. 1. It can be easily calculated by forward pass that.



Fig. 1. Flowchart of BPNN

$$x_s^{(i)} \approx a_L = pW_L W_{L-1} \cdots W_1 h_s^{(i)}.$$

It also means that

$$X \approx pW_L W_{L-1} \cdots W_1 H.$$

To enhance the clustering power of deep features, we combine our BPNN with two scatter quantities, namely intra-class scatter and total-class scatter. Our DDDNMF model is thereby created as follows:

$$L_{DDDNMF} = \frac{1}{2} \|X - pW_L W_{L-1} \cdots W_1 H\|_F^2 + \frac{\alpha}{2} \text{tr}(S_w) - \frac{\beta}{2} \text{tr}(S_t), \quad (2)$$

where S_w and S_t are respectively given by

$$S_w = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} (h_j^i - \bar{h}_i)(h_j^i - \bar{h}_i)^T,$$

$$S_t = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} (h_j^i - \bar{h})(h_j^i - \bar{h})^T.$$

They can be reformulated as the matrix forms shown below.

$$S_w = H \left(\frac{1}{n} I - \frac{1}{n} Z \right) H^T, \quad S_t = H \left(\frac{1}{n} I - \frac{1}{n^2} \mathbf{1}_{n \times n} \right) H^T,$$

where

$$Z = mdiag \left(\frac{1}{n_1} \mathbf{1}_{n_1 \times n_1}, \frac{1}{n_2} \mathbf{1}_{n_2 \times n_2}, \dots, \frac{1}{n_c} \mathbf{1}_{n_c \times n_c} \right).$$

The gradient descent method is exploited to deal with the optimization problem of minimizing the objective function (2) and the following DDNMF update rules are obtained:

$$\begin{aligned} W_i^{(t+1)} &= W_i^{(t)} \otimes \frac{\Phi_{L,i+1}^\top X H^\top \Phi_{i-1,1}^\top}{p^{(t)} \Phi_{L,i+1}^\top \Phi_{L,1} H H^\top \Phi_{i-1,1}^\top}, \\ H^{(t+1)} &= H^{(t)} \otimes \frac{p^{(t)} \Phi_{L,1}^{(t+1)\top} X + \frac{\alpha}{n} H Z + \frac{\beta}{n} H}{p^{(t)2} \Phi_{L,1}^{(t+1)\top} \Phi_{L,1}^{(t+1)} H + \frac{\alpha}{n} H + \frac{\beta}{n^2} H 1_{n \times n}}, \\ p^{(t+1)} &= \frac{\text{tr}(\Phi_{L,1}^{(t+1)} H^{(t+1)} X^\top)}{\text{tr}(\Phi_{L,1}^{(t+1)} H^{(t+1)} H^{(t+1)\top} \Phi_{L,1}^{(t+1)\top})}. \end{aligned} \quad (3)$$

where $\Phi_{k,l} = W_k W_{k-1} \dots W_l$.

The detailed derivation and convergence analysis on DDNMF will be reported in the next section.

4 Algorithm Discussion

This section focuses on deriving the iterative formula for H in (3) and discussing its convergence. The other update formulas in (3) can be deduced similarly and thus neglected here. To this end, we first compute the gradient of L_{DDNMF} with respect to H and obtain that

$$\begin{aligned} \nabla_H L_{DDNMF} &= -p \Phi_{L,1}^\top X - \frac{\alpha}{n} H Z - \frac{\beta}{n} H \\ &\quad + p^2 \Phi_{L,1}^\top \Phi_{L,1} H + \frac{\alpha}{n} H + \frac{\beta}{n^2} H 1_{n \times n}. \end{aligned} \quad (4)$$

Substituting (4) into the following gradient descent equation

$$H = H - \rho \nabla_H L_{DDNMF}$$

and letting the step length ρ be

$$\rho = \frac{H}{p^2 \Phi_{L,1}^\top \Phi_{L,1} H + \frac{\alpha}{n} H + \frac{\beta}{n^2} H 1_{n \times n}},$$

the update rule for H in (3) can be acquired by direct computation. The strategy of auxiliary function is utilized to show its convergence. The details are as follows.

Definition 1. $G(h, h')$ is an auxiliary function for $L(h)$, if for any vector $h, h' \in R^n$, there are $G(h, h') \geq L(h)$ and $G(h, h) = L(h)$.

Lemma 1. $L(h)$ is non-increasing with the iteration formula $h^{(t+1)} = \arg \min_h G(h, h^{(t)})$, if $G(h, h')$ is an auxiliary function for $L(h)$.

We just consider the objective function (2) with vector variable h_s as below:

$$L(h_s) = \frac{1}{2} \|x_s - p\Phi_{L,1}h_s\|_F^2 + \frac{\alpha}{2} \text{tr}(S_w) - \frac{\beta}{2} \text{tr}(S_t),$$

where $\Phi_{L,1} = W_L \cdots W_2 W_1$, x_s and h_s are the i th column of X and H , respectively. The auxiliary function $G(h_s, h_s^{(t)})$ for $L(h_s)$ is constructed by (5) in the following Theorem 1. The proof is omitted here for limited space.

Theorem 1. If $G(h_s, h_s^{(t)})$ is given by

$$\begin{aligned} G(h_s, h_s^{(t)}) &= L(h_s^{(t)}) + (h_s - h_s^{(t)})^\top \nabla L(h_s^{(t)}) \\ &\quad + \frac{1}{2} (h_s - h_s^{(t)})^\top D(h_s^{(t)}) (h_s - h_s^{(t)}), \end{aligned} \tag{5}$$

where $D(h_s)$ is a diagonal matrix with diagonal element

$$[D(h_s)]_{aa} = \frac{[p^2 \Phi_{L,1}^\top \Phi_{L,1} h_s + \frac{\beta}{n^2} H \mathbf{1}_{n \times 1} + \frac{\alpha}{n} h_s]_a}{[h_s]_a}$$

and $\mathbf{1}_{n \times 1}$ is an n -by-1 vector of ones, then $G(h_s, h_s^{(t)})$ is an auxiliary function for $L(h_s)$.

To find the stable point of $G(h_s, h_s^{(t)})$, we need to solve the extreme value problem $h_s^{(t+1)} = \arg \min_{h_s} G(h_s, h_s^{(t)})$. For this purpose, let $\nabla G(h_s) = 0$, namely,

$$\nabla G(h_s) = \nabla L(h_s^{(t)}) + D(h_s^{(t)}) (h_s - h_s^{(t)}) = 0.$$

Hence, we have

$$\begin{aligned} h_s^{(t+1)} &= h_s^{(t)} - D^{-1}(h_s^{(t)}) \nabla L(h_s^{(t)}) \\ &= h_s^{(t)} \otimes \frac{p^{(t)} \Phi_{L,1}^{(t+1)\top} x_s + \frac{\alpha}{n} H z_s + \frac{\beta}{n} h_s^{(t)}}{p^{(t)2} \Phi_{L,1}^{(t+1)\top} \Phi_{L,1}^{(t+1)} h_s^{(t)} + \frac{\alpha}{n} h_s^{(t)} + \frac{\beta}{n^2} H \mathbf{1}_{n \times 1}}. \end{aligned}$$

The update rule of H in (3) can be obtained by rewriting the above iteration into matrix form. It indicates from Lemma 1 that the objective function (2) is monotonic non-increasing under the update rule of H in (3).

5 Experiment

To assess the proposed DDNMF algorithm, this section will conduct clustering experiments using human facial images. Four related algorithms, including NMF [1], MNMF [9], RDNBMF [4] and BPDNMF [8], are selected for comparison. In NMF, the number of basis images is set to 300. We perform three-times decomposition in MNMF with

dimensions of 400, 300, and 300 respectively. The layer number in RDNBMF is 2 as given in [4] and the dimensions are 400 and 300 respectively. In BPDNMF, the neurons number in the first layer is the same as the number of training samples, and the number of neurons is 400 and 500 respectively in the next two layers. While for DDNMF, we create a three-layer neural network architecture, the number of neurons in each layer is 400, 500, 600, respectively. The penalty parameters α and β are assigned to 0.01 and 0.001, respectively. We use average clustering accuracy (ACC) and normalized mutual information (NMI) to evaluate the clustering effectiveness. The high values of ACC and NMI mean good clustering performance.

5.1 Facial Image Databases

The ORL database comprises of 400 grayscale facial images of 40 distinct persons. Each subject has 10 different images. The images change in pose and expression. The resolution of each image is 30×25 . Figure 2 shows 10 images of one individual from ORL database.



Fig. 2. Images of one person from ORL database

While for Jaffe database, we choose 200 facial images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. The resolution of each image is 256×256 . Part images of one person from Jaffe database are shown in Fig. 3.



Fig. 3. Images of one person from Jaffe database

5.2 Results on ORL Database

We randomly select 6 images from each individual for training while the rest images are for clustering. The number of clusters (CN) varies from 8 to 40 with a gap of 4. The clustering results are respectively recorded in Table 1 (ACC) and Table 2 (NMI).

It can be seen that all deep NMF methods except MNMF outperform the single-layer NMF method. The NN-based deep NMF methods, such as BPDNMF and DDNMF, have better clustering results than single-layer NMF-based deep NMF methods, such as MNMF and RDNBMF. This is because the accumulated error caused by multiple decompositions makes their overall decomposition error larger when single-layer NMF is repeatedly utilized to generate the deep hierarchical structure. In particular, MNMF method is even inferior to the single-layer NMF method. Among them, our DDNMF has the best clustering performance on the ORL database due to the use of class label information and neural networks for model building.

Table 1. Mean Accuracy (%) versus Clustering Numbers (CN) on ORL database.

CN	8	12	16	20	24	28	32	36	40
NMF [1]	50.93	52.91	51.09	53.50	51.98	49.02	48.98	45.90	46.25
MNMF [9]	38.43	35.83	32.03	32.75	30.72	29.55	29.60	29.37	29.43
RDNBMF [4]	57.81	55.41	53.75	56.75	51.35	49.91	50.85	49.23	49.12
BPDNMF [8]	62.81	61.87	59.84	57.25	55.41	50.71	52.57	49.65	49.81
DDNMF	70.31	68.95	61.41	66.13	61.77	59.38	62.03	58.82	58.93

Table 2. Mean NMI (%) versus Clustering Numbers (CN) on ORL database.

CN	8	12	16	20	24	28	32	36	40
NMF [1]	60.04	68.05	68.87	71.93	72.81	71.48	72.17	70.73	71.34
MNMF [9]	41.92	47.81	50.37	54.64	56.02	56.50	58.92	59.79	61.33
RDNBMF [4]	65.48	70.01	70.37	73.45	72.79	72.71	74.15	73.09	73.99
BPDNMF [8]	71.57	74.51	74.03	74.90	75.15	72.79	75.26	74.68	74.82
DDNMF	79.45	81.80	78.35	81.28	80.40	80.19	81.86	81.44	80.80

5.3 Results on Jaffe Database

On the Jaffe database, 15 randomly selected images from each class are used for training, and the rest of the images are for clustering. The number of selected clusters ranges from 2 to 10. The clustering results are tabulated in Table 3 (ACC) and Table 4 (NMI), respectively. It demonstrates that the proposed DDNMF achieves the best clustering performance among all the compared algorithms.

Moreover, except for MNMF, all other deep NMF methods surpass the shallow NMF method in most cases. In contrast, MNMF has better computational efficiency, but it simply uses NMF recursively to generate the deep structure, and thus cannot obtain entire decomposition with high precision. This causes its clustering performance to be not as good as the single-layer NMF algorithm.

Table 3. Mean Accuracy (%) versus Clustering Numbers (CN) on Jaffe database.

CN	2	4	6	8	10
NMF [1]	78.00	69.00	61.00	59.75	59.00
MNMF [9]	67.00	43.20	36.67	33.50	34.20
RDNBMF [4]	79.00	62.50	70.00	63.00	62.20
BPDNMF [8]	85.00	70.00	68.00	56.75	53.80
DDNMF	99.00	82.50	75.00	68.75	64.00

Table 4. Mean NMI (%) versus Clustering Numbers (CN) on Jaffe database.

CN	2	4	6	8	10
NMF [1]	42.98	66.09	65.92	67.47	69.64
MNMF [9]	18.53	23.57	29.87	33.81	39.62
RDNBMF [4]	42.57	63.00	75.93	71.94	75.71
BPDNMF [8]	61.25	69.00	68.97	62.39	63.46
DDNMF	99.00	81.92	80.22	76.55	77.73

6 Conclusion

In this paper, we propose a novel discriminant deep NMF approach based on a neural network. The DDNMF model is created by incorporating the supervised regularization into the loss function of a special neural network. We solve the optimization problem of the DDNMF model using the GDM method and develop the DDNMF algorithm, which is shown to be convergent. Our DDNMF algorithm has achieved superior performance in image clustering. The research results also conclude that NN-based deep NMF approaches are better than NMF- based deep NMF approaches in most cases. Furthermore, combining neural networks with labeled information will further enhance the clustering effect of the deep NMF approaches.

Acknowledgements. This work was partially supported by the Natural Science Foundation of Shenzhen (20200815000520001) and the Interdisciplinary Innovation Team of Shenzhen University. We would like to thank the Olivetti Research Laboratory and Kyushu University for contributions of ORL database and Jaffe database, respectively.

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791 (1999)
2. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* **13** (2000)

3. Zhu, W., Yan, Y.: Label and orthogonality regularized non-negative matrix factorization for image classification. *Sig. Process. Image Commun.* **62**, 139–148 (2018)
4. Zhao, Y., Wang, H., Pei, J.: Deep non-negative matrix factorization architecture based on underlying basis images learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(6), 1897–1913 (2019)
5. Chen, W.S., Ge, X., Pan, B.: A novel general kernel-based non-negative matrix factorisation approach for face recognition. *Connect. Sci.* **34**(1), 785–810 (2022)
6. Deng, P., Li, T., Wang, H., Wang, D., Horng, S.J., Liu, R.: Graph regularized sparse non-negative matrix factorization for clustering. *IEEE Trans. Comput. Soc. Syst.* (2022). <https://doi.org/10.1109/TCSS.2022.3154030>
7. Li, X., Cui, G., Dong, Y.: Graph regularized non-negative low-rank matrix factorization for image clustering. *IEEE Trans. Cybern.* **47**(11), 3840–3853 (2016)
8. Zeng, Q., Chen, W.S., Pan, B.: Bp neural network-based deep non-negative matrix factorization for image clustering. In: International Conference on Intelligent Computing, pp. 378–387. Springer (2020)
9. Ahn, J.H., Kim, S., Oh, J.H., Choi, S.: Multiple nonnegative-matrix factorization of dynamic pet images. In: Proceedings of Asian Conference on Computer Vision, pp. 1009–1013. Citeseer (2004)
10. Cichocki, A., Zdunek, R.: Multilayer nonnegative matrix factorisation. *Electron. Lett.* **42**(16), 947–948 (2006)
11. Yuan, Y., Zhang, Z., Liu, G.: A novel hyperspectral unmixing model based on multilayer nmf with hoyer's projection. *Neurocomputing* **440**, 145–158 (2021)
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back- propagating errors. *Nature* **323**(6088), 533–536 (1986)



A Feature Extraction Algorithm for Enhancing Graphical Local Adaptive Threshold

Shaoshao Wang, Aihua Zhang^(✉), and Han Wang

College of Control Science and Engineering, Bohai University, Jinzhou 121013, Liaoning, China
Jsxinxixi_zah@163.com

Abstract. In order to solve the problem that the ORB algorithm increases the probability of feature point loss and mis-matching in some cases such as insufficient light intensity, low texture, large camera rotation, etc. This paper introduces an enhanced graphical local adaptive thresholding (EGLAT) feature extraction algorithm, which enhances the front-end real-time input image to make the blurred texture and corners clearer, replacing the existing ORB extraction method based on static thresholding, the local adaptive thresholding algorithm makes the extraction of feature points more uniform and good quality, avoiding the problems of over-concentration of feature points and partial information loss. Comparing the proposed algorithm with ORB-SLAM2 in a public dataset and a real environment, the results show that our proposed method outperforms the ORB-SLAM2 algorithm in terms of the number of extracted feature points, the correct matching rate and the matching time, especially the matching rate of feature points is improved by 18.7% and the trajectory error of the camera is reduced by 16.5%.

Keywords: SLAM · Adaptive algorithm · Image matching · Feature extraction · Image processing

1 Introduction

Simultaneous Localization and Mapping (SLAM) technology has been used in intelligent robots, unmanned vehicles, human-computer interaction, etc. With the continuous progress and development of image processing technology, more and more researchers have applied vision processing to SLAM systems [1]. Extraction and matching of feature points is an important part of visual SLAM for environment perception, which plays a very important role in target recognition, construction of the map and back-end optimization. Harris et al. [2] proposed Harris corner and edge detection algorithm, which mainly uses differential operation and autocorrelation matrix for corner detection. David Lowe [3, 4] proposed the scale Invariant Feature Transform (SIFT) algorithm, which is the most stable point detection algorithm with high accuracy. Herbert Bay [5] improved the SIFT algorithm and realized SURF algorithm, which is three times faster than SIFT algorithm, However, compared with the ORB algorithm [6] proposed by Raul Mur-Artal et al., It has better effect than SIFT and SURF in terms of real-time and extraction speed of feature point extraction.

Based on the ORB algorithm, Raul Mur-Artal team proposed the ORB-SLAM2 [7, 8] system, which can not only support monocular, binocular and RGB cameras at the same time, but also calculate the pose of cameras in real time and construct accurate maps, which is being used by more and more developers with excellent algorithm framework. However, since the algorithm generates a sparse 3D map of the scene, more feature points need to be extracted to ensure the accuracy of the map. Therefore, in order to obtain more feature points of quality. Literature [9] proposed the SURB algorithm combining ORB and SURF algorithm for feature point detection, but this method needs to constantly count the Harr's wavelet feature directions in the circular neighborhood of feature points to determine the main direction, so feature point extraction is very time-consuming. Fan [10, 11] proposed a pixel intensity based local variance adaptive threshold extraction method, but in the case of weak light intensity will lead to feature points extraction effect is very bad, this paper [12] proposes a K - means clustering is used to calculate the threshold image entropy method, algorithm to calculate the information entropy of image block less need to remove part of the information of image block, As a result, some scenes have no feature points to represent information, which reduces the accuracy of maps. Ding et al. [13] used THE KSW entropy method to calculate the FAST threshold, but the algorithm was susceptible to the influence of many moving scenes and depended on the initial static threshold. Based on the shortcomings of the above algorithms, this paper proposes an algorithm of enhanced Graphical Local Adaptive Threshold (EILAT) feature point extraction, which solves the problem of feature point extraction loss in the case of weak texture in [10, 11] and improves the time of feature point extraction [8][9]. Adaptive threshold algorithm is used to overcome the disadvantages of relying on initial static threshold and information loss in [11] and [13], and improve the speed and quality of feature point extraction. The main contributions of this paper are as follows:

- (1) In order to solve the problems of missing information and many matching error points in the extraction part of ORB feature points, the ORB-SLAM2 algorithm is improved and proposed a feature point extraction algorithm with enhanced graphical local adaptive threshold (EGLAT-SLAM).
- (2) Image enhancement processes some pictures with insufficient lighting or unclear texture, making the texture of the image clearer, which is beneficial to the subsequent feature extraction, and solves the situation of partial information loss of image feature points and insufficient information tracking loss.
- (3) Using the locally adaptive threshold algorithm to improve the shortcomings of the ORB-SLAM2 fixed threshold, making the acquisition of the image data has more quality of feature points are screened out and improved the quality of the extraction, characteristics of the phenomenon of mismatch are greatly improved, ensuring the validity of the data transmission system and the back-end data processing speed, which map is more accurate.

Based on the above improvements, the number of feature points extracted, the number of feature matching points and the error of the trajectory generated are compared between the EGLAT-SLAM and ORB-SLAM2 by using tum data set and real environment. The results showed that EGLAT-SLAM algorithm has better quality of feature point

extraction and better matching results than ORB-SLAM2 algorithm, the trajectory error is reduced by 16.5% compared with ORB-SLAM2.

2 Relate Work

2.1 ORB Feature Extraction

Extraction of ORB features is composed of two parts, improved FAST corner point and binary descriptor BRIEF [14] [15], based on the construction of image pyramid in different environments feature point extraction has the advantages of fast and robust, ORB feature extraction process is divided into two steps, firstly, FAST is a corner point, which mainly detects the part of local pixel with obvious grayscale change and is known for FAST is known for its speed, when a pixel is different from the pixel in the field is detected this point may be a corner point, so in the image feature extraction only need to compare the size of the pixel brightness, Specific detection process (see Fig. 1), Secondly, computes the feature point descriptors.

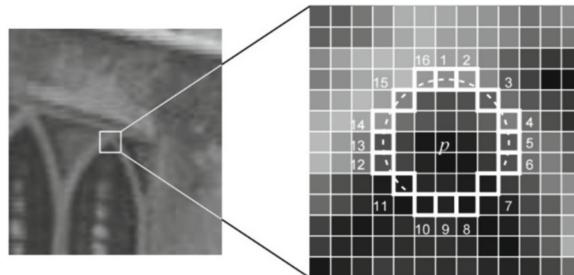


Fig. 1. FAST feature point detection

ORB feature extraction adds descriptors of scale and rotation, invariance of scale is achieved by constructing a pyramid, extracted corners is performed on each layer of the pyramid and feature points are equally distributed to each layer of the pyramid according to the area, set the length of the input image be L , the width of the input image be denoted as W , and the scaling factor be s . then the total area of the whole pyramid is:

$$S = L \times W \times \frac{1 - [s^2]^n}{1 - [s^2]} = C \frac{1 - [s^2]^n}{1 - [s^2]} \quad 0 < s < 1 \quad (1)$$

For layer the number of feature points that should be assigned is:

$$N_a = \frac{N \{1 - [s^2]\}}{\{1 - [s^2]^n\}} [s^2]^a \quad (2)$$

Rotation invariance is realized by the gray centroid method, which can be understood as taking the gray value of a pixel as the center of weight in the image area, the vector

whose centroid points to the centroid as the main direction of the key point and the image block moment is defined as:

$$m_{pq} = \sum_{x,y} x^p y^q I(x, y) \quad (3)$$

The quality of the image block can be found by the moments as:

$$C = (c_x, c_y) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right) \quad (4)$$

The principal direction of the key point can be expressed that the direction vector from the center of the circular image shape O to the center of mass C, therefore, the direction angle of the feature point is expressed as:

$$\theta = \arctan(m_{01} / m_{10}) \quad (5)$$

BRIEF is a binary descriptor that consists of many zeros and ones, these zeros and ones encode and represent the size relationship of two random pixels (such as P and Q) near the key point. The speed of this calculation is very fast because of the randomly selected points, which is suitable for real-time image extraction and matching.

2.2 CLAHE Algorithm

In the current environment, there are always blurred images and low texture during the acquisition process, because of factors as the camera moving too fast or insufficient lighting. How to better process blurred images is particularly important. Adaptive histogram equalization [16] (AHE) is a computer image processing technology, which is used to improve image contrast. This method overamplifies noise in the relatively uniform area of the image, resulting in poor texture enhancement effect of the image. Therefore, the method of finite contrast adaptive histogram equalization [17] (CLAHE) is proposed to limit noise amplification and local enhancement by limiting the height of the local histogram. This method divides the image into multiple sub-areas, such as square sliding window with M*M size, and the local mapping function is expressed as:

$$S = \frac{d(m(i))}{di} = H_{ist}(i) \times \frac{255}{M \times M} \quad (6)$$

The histogram of each sub region is classified, and each sub region is histogram equalized respectively, then cut the histogram to make its amplitude lower than a certain upper limit T, the maximum height of histogram is:

$$H_{\max} = S_{\max} \times \frac{255}{M \times M} \quad (7)$$

The clipping part is evenly distributed over the whole gray range to ensure that the histogram remains unchanged with the total area (see Fig. 2).

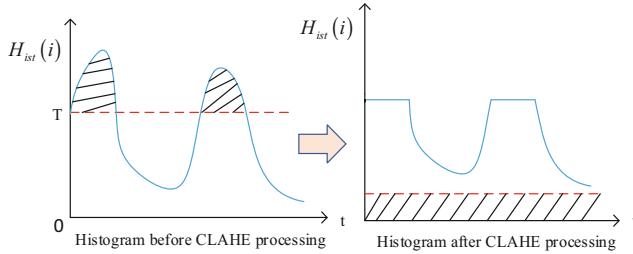


Fig. 2. The algorithm clipping keeps the total area of the histogram unchanged

The transformed gray value is obtained by interpolating each pixel, the histogram of the enhanced image is equalized and finally improved histogram expression is:

$$H_{ist}(i) = \begin{cases} H_{ist}(i) + L H_{ist}(i) < T \\ H_{max} H_{ist}(i) \geq T \end{cases} \quad (8)$$

where L represent area of histogram clipping. The transformation function is exactly as originally defined for the central pixel of the block. The other pixels are obtained by interpolating the transformation functions of those four blocks that are close to them. Varying the maximum slope of the mapping function S_{max} and the height of the corresponding maximum histogram H_{max} , which obtained images with different enhancement's effects.

2.3 Local Adaptive Thresholding Algorithm

The general purpose of image thresholding is to separate the target and background regions from grayscale images, the maximum interclass variance [18] was proposed an adaptive thresholding determination method by a Japanese scholar (Nobuyuki Otsu), what is difficult for this algorithm to achieve desiring feature extraction effect by setting a global fixed threshold [15]. However, this method is difficult to achieve the ideal feature extraction effect, therefore, an adaptive local threshold method is proposed, which calculated the local threshold which each image according to the brightness distribution of different regions of the image, this method ensures that the threshold of each pixel in the image will change with the change of its surrounding pixels. The specific idea is calculated image with (x,y) , the standard deviation in an image, and the mean value of the set of pixels in the domain S_{max} centered at (x,y) are computed, the expression of the local threshold is:

$$T_{xy} = a\delta_{xy} + b m_{xy} \quad (9)$$

where a, b are non-negative numbers, the calculated thresholds are returned to the feature extraction algorithm to achieve local adaptive threshold extraction.

3 System Overview

3.1 Algorithm Flow

In this paper, it is proposed that an ORB feature point detection algorithm with enhanced graphical local adaptive threshold, The whole system framework (Fig. 3(b)) is shown

below. The histogram of the input image is defined as $H_{ist}(i)$, the histogram of the original image is normalized and the sum of the group distances is 255, The integral sum of the calculated histogram is expressed as:

$$H'_{ist} = \sum_{0 \leq j \leq i} H'_{ist}(j) \quad (10)$$

The local mapping function S is used to process the gray level of each image and a gray scale mapping table $S(i)$ is obtained, finally, each pixel in the original image is modified according to the corresponding gray value in table $S(i)$, The histogram distribution of the image is more uniform through CLAHE's processing, the processed image is defined as $I(x, y)$ and the Gaussian smoothing filter is $G_\sigma(x, y)$, the image noise is reduced through convolution as shown in:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (11)$$

$$L(x, y) = G_\sigma(x, y) * I(x, y) \quad (12)$$

where $*$, $I(x, y)$, $G_\sigma(x, y)$ are the convolution operator.

The collected image to build the image pyramid [19]. Let P be a feature point and $R(i, j)$ represent a locally adjustable template, the grayscale difference between different adjacent pixels is calculated which is defined as:

$$\Delta\delta = |L(x_1, y_1) - L(x_2, y_2)| \quad (13)$$

where $L(x_1, y_1)$ and $L(x_2, y_2)$ respectively represent the gray levels of adjacent pixels.

Calculate the gray values of all adjacent pixels $\Delta\delta_i$ and the mean values of pixels m_{xy} as shown in:

$$\Delta\delta_i = \sum_{0 \leq j \leq i} \Delta\delta_j \quad (14)$$

$$m_{xy} = \frac{\Delta\delta_i}{i} \quad (15)$$

Then the standard deviation δ_{xy} and local threshold of the pixel T_{xy} set are calculated, which can be expressed as:

$$\delta_{xy} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\delta_i - m_{xy})^2} \quad (16)$$

Substitute the calculated δ_{xy} and m_{xy} into formula (4) to obtain the local adaptive threshold T_{xy} , then continuously calculate the feature points of the whole image according to the threshold.

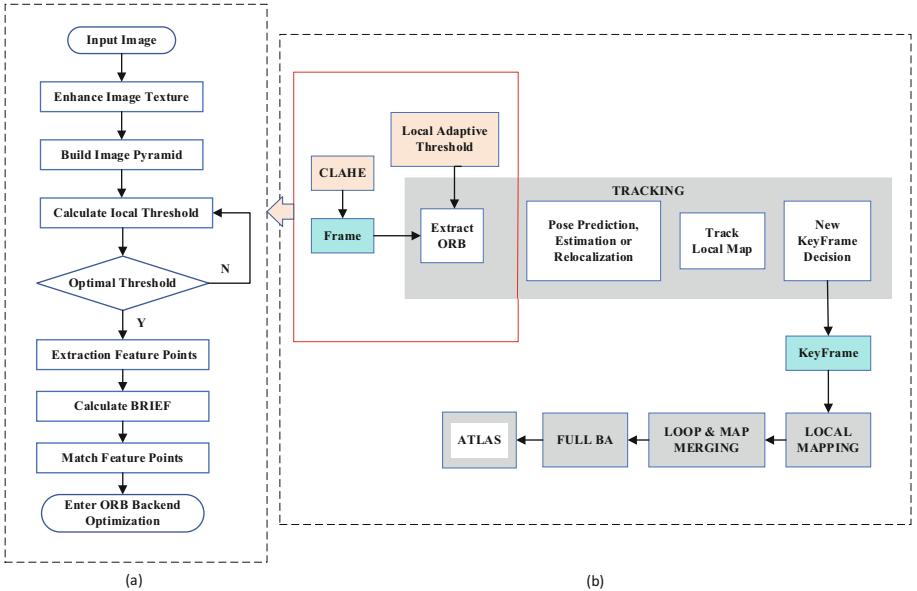


Fig. 3. The overall structure and algorithm flow of the system

The number of pyramid layers can be set according to the needs of the environment, the number of extracted feature points is distributed on the image pyramid layers according to the rules, and then according to the predefined size of the image block the division is performed, the threshold corresponding to each image block is compared layer by layer to see if it is optimal. If the threshold value of the image block is not locally optimal, the optimal threshold is input into the quadtree algorithm to extract the optimal feature points, as shown in Fig. 3(a), the optimal feature points and descriptors are extracted for feature matching, then the matching information is transmitted to the back-end processing of the ORB-SLAM system.

3.2 Feature Point Matching

Feature matching is a key step in the visual SLAM, on the one hand, matching mechanism can solve the problem of SLAM associated data, such as the current see road signs and saw before sign the corresponding relationship between, on the other hand, matching can figure with figure or image descriptor precise matching between map, the follow-up of robot's pose estimation and map data optimization to reduce operation. Bag of words (BOW) used in the ORB-SLAM2 system which mainly utilizes the Feature Vector in BoW to accelerate feature matching [20].

In the actual feature matching, BoW [21] is generated online, a certain number of feature points and descriptors are transmitted to make the dimensionality consistent with that in the vocabulary tree, and then for each descriptor of a feature point, the descriptor from the vocabulary tree created offline starts to find its own position, starting from the root node, the Hamming distance is calculated with the descriptor of each node, the one

with the smallest Hamming distance is selected as the node where it is located, traversing all the way to the leaf node, and the branch clustering structure (see Fig. 4), the black line represents the process of a feature point from the root node to the leaf node.

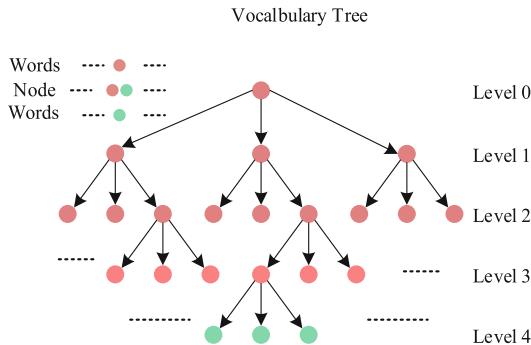


Fig. 4. Tree cluster structure.

4 Experimental Results

4.1 Experimental Environment

All data in this paper is obtained from a laptop and an external camera processing, the laptop operating system is Ubuntu 20.04, the processor Intel (R) Core (TM) i7-7700U CPU @ 3.6 GHz, the running environment is Visual Studio Code and opencv3.4.15, the program is written in C++, using the data set the performance of the proposed method is analyzed using TUM and field environment.

4.2 Analysis of Experimental Results

Firstly, the initial image processing was carried out, the experiment selected the data set tum and three groups of pictures in the real environment for comparison, as shown in Fig. 5, the image enhancement technology of CLAHE is used to enhance the texture of three groups of images with insufficient daylighting (Fig. 6).

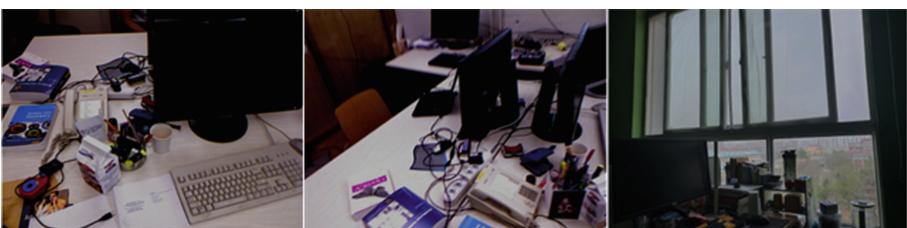


Fig. 5. An original image that has not been enhanced.



Fig. 6. Image enhanced by CLAHE algorithm.

After CLAHE processing, it is obvious that some edges and corners with insufficient lighting and low texture in the original graphics have been enhanced, and the lines of the image are clearer, which leads to easier feature point extraction in the next step, has good robustness, and makes up for the loss of feature point extraction caused by the unclear image gradient.

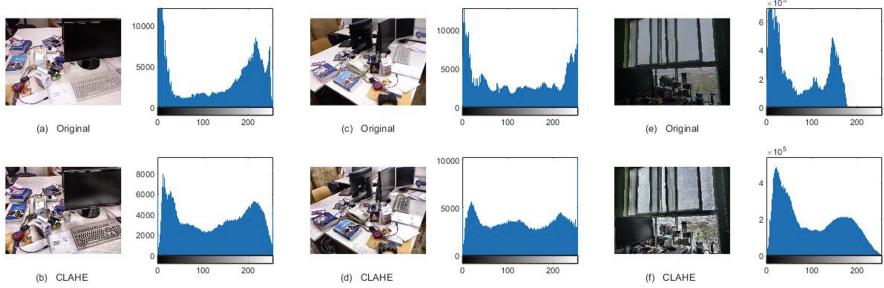


Fig. 7. Histogram before and after CLAHE algorithm enhancement.

The histogram of the image is used to fully represent the number of pixels at each brightness level of the image, showing the distribution of pixels in the image, with the horizontal axis representing brightness values from 0 to 255. The vertical axis represents the number of pixels corresponding to brightness in the photo. The two groups of images in Fig. 7 shows the construction of histograms. Figure 7(a)(c)(e) and Fig. 7(b)(d)(f) respectively correspond to the pixel distribution before and after image enhancement. The results show that the histogram distribution of the image processed by the enhancement algorithm is more uniform.

4.3 Extraction and Matching After Local Adaptive Threshold

After image enhancement processing, EGLAT-SLAM algorithm is used to extract feature points and calculate descriptors. The processing of local adaptive threshold makes the extraction of image feature points more uniform, so that more high-quality feature points are extracted for feature matching. In order to verify that the performance of this method is better than ORB-SLAM2, the images of data set and real environment are tested respectively, Fig. 8 is the result of feature point extraction of four groups of pictures.

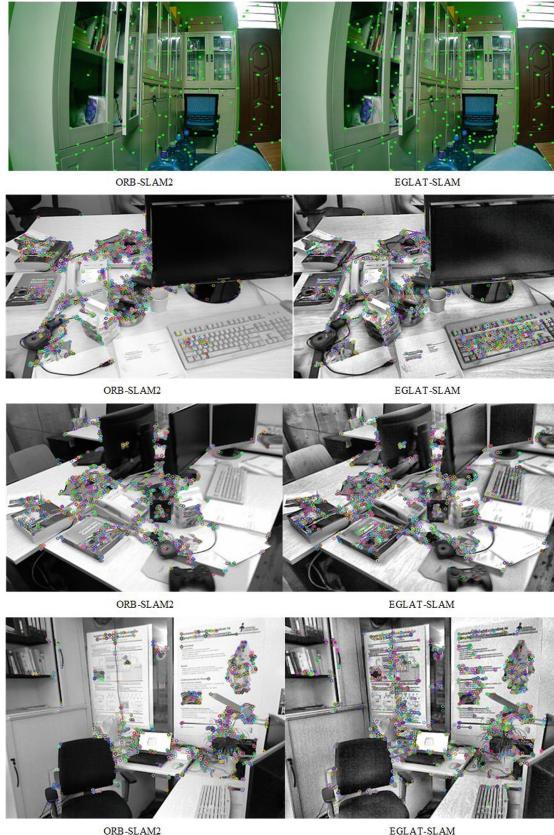


Fig. 8. Feature point extraction results of EGLLAT-SLAM algorithm

It can be seen from the above feature point extraction results that, on the one hand, the number of feature points extracted by ORB-SLAM2 algorithm is very small due to the existence of low texture edges and corners, on the other hand, some quality points are lost due to the defect of fixed threshold of the extraction algorithm. This phenomenon is more prominent in the feature point extraction of computer keyboard in the first and second groups of pictures. However, after EGLLAT-SLAM algorithm processing, the extraction of feature points of computer keyboard has been significantly improved, in the third group of experimental images in real environment, what is obvious is that the number of feature points extracted by our algorithm is more and more uniform, which fully verifies the effectiveness of our method.

As Fig. 9 shows the result of feature point matching, what is obvious is that EGLLAT-SLAM is better than ORB-SLAM2 in feature matching after the enhancement of the same image, because our algorithm makes the local image constantly adjust the selection threshold, so that more high-quality feature points can be extracted for matching, which improves the situation of tracking loss of low texture image data during the operation of ORB-SLAM2.

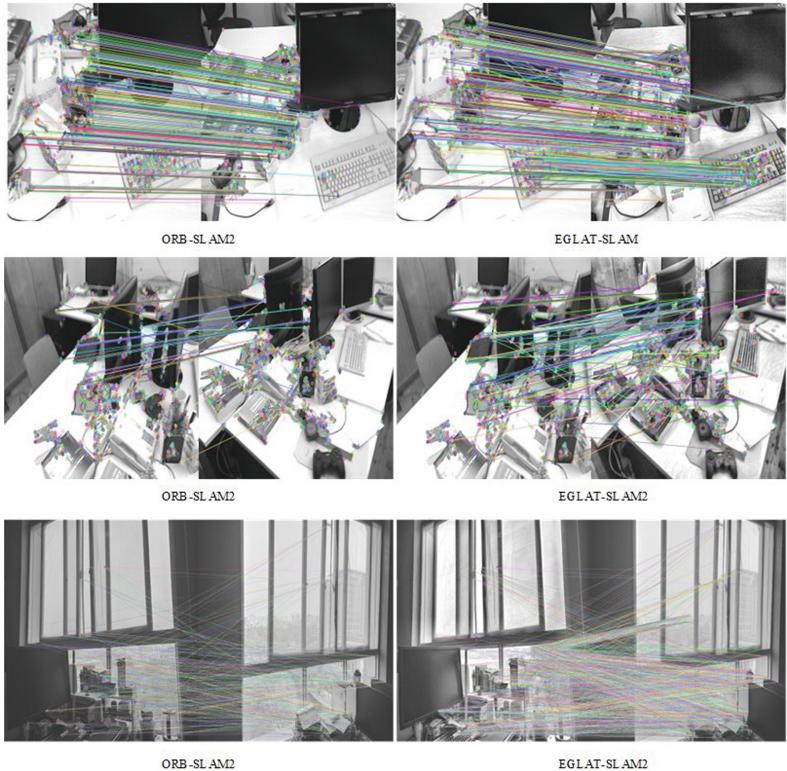


Fig. 9. The result of feature point matching of EGLAT-SLAM algorithm.

In other sequences of the data set tum, the matching time (s) of the algorithm and the number of mismatched feature points are tested, the results are shown in Table 1.

Table 1. The time of algorithm matching and the number of false matching points

	Match			Error match	
	ORB-SLAM2	EGLAT-SLAM	Improved	ORB-SLAM2	EGLAT-SLAM
Fre1_xyz	0.00746	0.00504	24.2%	25/358	16/358
Fre1_rpy	0.00928	0.00698	23%	31/502	19/502
Fr1_desk	0.01421	0.01133	29%	67/1049	43/1049
Fr3_walking_xyz	0.00545	0.00425	12%	36/638	25/638
Fr3_walking_half	0.01242	0.01191	5%	64/1079	47/1079
Mean	0.00976	0.00789	18.6%	223/3626	150/3626

It can be concluded that the matching time of EGLAT-SLAM algorithm is 18.7% higher than that of ORB-SLAM2 algorithm, and the number of mismatched feature points is reduced by 73, which is 2.01% higher overall.

4.4 Trajectory Error Evaluation

The superior SLAM algorithm should not only improve the matching efficiency, but also take advantage of the real-time map building and accurate estimation of camera poses in the later stage. Therefore, the method in this paper is integrated into the whole ORB-SLAM2 algorithm process, the error of the generated map trajectory is evaluated on the data set and the Fre1_xyz sequence is used to generate the error map of the trajectory, as shown in Fig. 10.

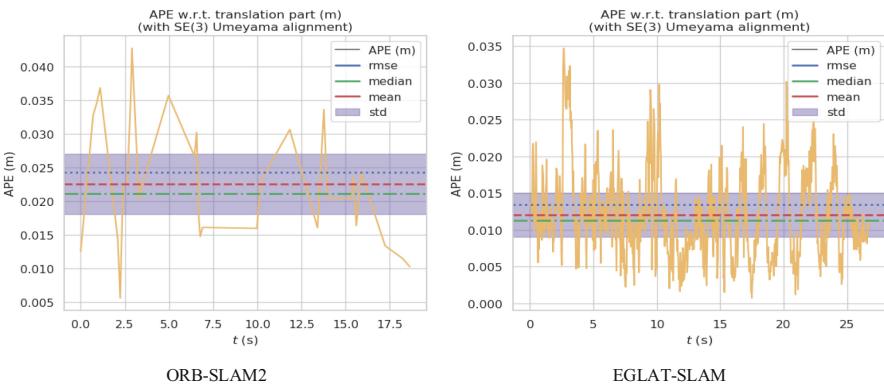


Fig. 10. Track error evaluation and comparison results.

Similarly, other sequences in the dataset were tested accordingly, as shown in Table 2. From the results of the sequence tests of each data, which can be seen that EGLAT-SLAM generates actual trajectories with less error than ORB-SLAM2, Due to the depth of the scene information.

Table 2. Data set trajectory error

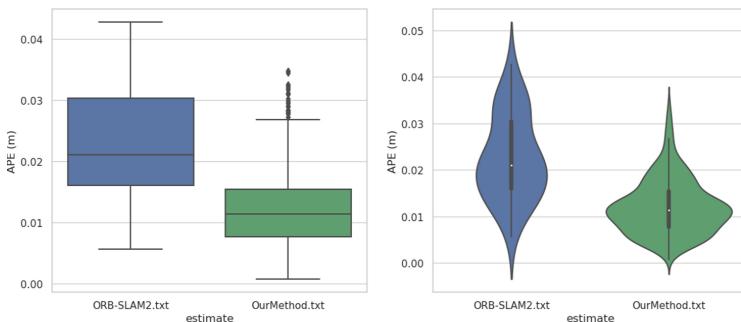
	RMSE		Mean	
	ORB-SLAM2	EGLAT-SLAM	ORB-SLAM2	EGLAT-SLAM
Fre1_xyz	0.0243	0.0134	0.0225	0.0128
Fre1_rpy	0.0324	0.0305	0.0265	0.0183
Fre1/360	0.0284	0.0291	0.0268	0.0271
Fre1/floor	0.0194	0.0185	0.0182	0.0179

(continued)

Table 2. (continued)

	RMSE		Mean	
	ORB-SLAM2	EGLAT-SLAM	ORB-SLAM2	EGLAT-SLAM
Fre1/desk	0.0165	0.0154	0.0157	0.0148
Fre2_xyz	0.0098	0.0087	0.0092	0.0085
Fre2_rpy	0.0146	0.0137	0.0142	0.0131
Fr2/360_kidnap	0.0095	0.0107	0.0087	0.0101
Fr2/desk	0.0079	0.0068	0.0075	0.0061
Fr2/360_hemisphere	0.0214	0.0211	0.0205	0.0198
Fr2/360_kidnap	0.0098	0.0089	0.0094	0.0091
Fr3/walking_rpy	0.0051	0.0048	0.0046	0.0041
Fr3/walking_xyz	0.2645	0.2358	0.1878	0.1376
Fr3/walking_half	0.4589	0.3187	0.3142	0.2568

The maximum absolute positional error (APE) of ORB-SLAM2 peaked at 0.042, while that of EGLAT-SLAM was only 0.035. EGLAT-SLAM reduced the maximum APE trajectory peak by 16% compared with that of ORB-SLAM2. The performance of the algorithm in the data set is shown in Fig. 11. From the results of graphic display, we can see that the performance of EGLAT-SLAM algorithm in generating the trajectory of the map is better than ORB-SLAM2.

**Fig. 11.** Error performance diagram of trajectory of algorithm.

5 Conclusion

In this paper, this paper presents a novel method for extracting ORB feature points based on local adaptive thresholding after image enhancement, the algorithm solves the problem of losing feature points extracted from low-texture images in the ORB

system and makes the speed of feature point extraction and matching in the SLAM system significantly improved. The algorithm is tested using public datasets and realistic environments, the results show that the EGLAT-SLAM algorithm outperforms the ORB-SLAM2 in terms of processing time of images, number of feature points extracted, correct matching rate and trajectory error of maps. In the future, we will further improve performance of this algorithm in different situations of the environment, which is ported to the vision processing system of the Kentucky robot.

References

1. Ahn, H.S., Sa, I., Choi, J.Y.: PDA-based mobile robot system with remote monitoring for home environment. *IEEE Trans. Consum. Electron.* **55**(3), 1487–1495 (2009)
2. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of Alvey Vision Conference, Manchester, U.K., pp. 147–151 (1988)
3. Lowe, D.G.: ‘Distinctive image features from scale-invariant keypoints?’ *Int. J. Comput. Vis.* **2**(60), 91–110 (2004)
4. Ng, D.P.C., Henikoff, S.: SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**(13), 3812–3814 (2003)
5. Sheng, H., Wei, S., Yu, X., Tang, L.: Research on binocular visual system of robotic arm based on improved SURF algorithm. *IEEE Sensors J.* **20**(20), 11849–11855 (2020)
6. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 2564–2571 (2011)
7. Mur-Artal, R., Tardós, J.D.: ‘ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras.’ *IEEE Trans. Robot.* **33**(5), 1255–1262 (2017)
8. Sun, K.: ‘Research on image matching and scene 3D reconstruction. *Huazhong Univ. Sci. Technol.* **10**(10), 13–22 (2017)
9. Wang, X., Zou, J., Shi, D.: An improved ORB image feature matching algorithm based on SURF. In: 2018 3rd International Conference on Robotics and Automation Engineering (ICRAE), Guangzhou, China, pp. 218–222 (2018)
10. Fan, G.: Research on visual SLAM algorithm of mobile robot in dynamic indoor scene. M.S. thesis, Xi'an Univ. Technol., Xi'an, China (2020)
11. Wu, R., Pike, M., Lee, B.G.: DT-SLAM: dynamic thresholding based corner point extraction in SLAM system. *IEEE Access* **9**, 91723–91729 (2021)
12. Ma, Y., Shi, L.: A modified multiple self-adaptive thresholds fast feature points extraction algorithm based on image gray clustering. In: Proceedings of International Applied Computational Electromagnetics Society Symposium (ACES), pp. 1–5 (2017)
13. Sun, C., Wu, X., Sun, J., Qiao, N., Sun, C.: Multi-stage refinement feature matching using adaptive ORB features for robotic vision navigation. *IEEE Sens. J.* **22**(3), 2603–2617 (2022)
14. Xu, J., Chang, H.-w., Yang, S., Wang, M.: Fast feature-based video stabilization without accumulative global motion estimation. *IEEE Trans. Consumer Electron.* **58**(3), 993–999 (2012)
15. Yin, D., et al.: A feature points extraction algorithm based on adaptive information entropy. *IEEE Access* **8**, 127134–127141 (2020)
16. Sino, H.W., Indrabayu, Areni, I.S.: Face recognition of low-resolution video using gabor filter & adaptive histogram equalization. In: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, pp. 417–421 (2019)

17. Zhou, M., Jin, K., Wang, S., Ye, J., Qian, D.: Color retinal image enhancement based on luminosity and contrast adjustment. *IEEE Trans. Biomed. Eng.* **65**(3), 521–527 (2018)
18. Wang, L.-H., et al.: Automated classification model with OTSU and CNN method for premature ventricular contraction detection. *IEEE Access* **9**, 156581–156591 (2021)
19. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Trans. Rob.* **37**(6), 1874–1890 (2021)
20. Sinaga, K.P., Yang, M.: Unsupervised K-Means clustering algorithm. *IEEE Access* **8**, 80716–80727 (2020)
21. Guo, S., Guo, W.: Process monitoring and fault prediction in multivariate time series using bag-of-words. *IEEE Trans. Autom. Sci. Eng.* **19**(1), 230–242 (2022)



Person Re-identification Based on Transform Algorithm

Lei Xie¹, Chao Wang¹(✉), Xiaoyong Yu¹, Aihua Zheng², and Guolong Chen³

¹ School of Informatics and Engineering, Suzhou University, Suzhou 234000,
People's Republic of China
szxycw@126.com

² Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui
University, Hefei 230601, People's Republic of China

³ School of Computer Science and Information Engineering, Bengbu University,
Bengbu 233000, People's Republic of China

Abstract. Person Re-identification (Person ReID) is a new technology emerging in the field of intelligent video analysis in recent years, aiming to solve the problem of person re-identification and retrieval under cross-lenses and scenes. It is also a hot research topic of computer vision in recent years. However, the study of Person ReID faces many challenges, such as low image resolution, visual angle change, attitude change, light change and occlusion. Research on target ReID has focused on Person ReID and vehicle ReID, and most state-of-the-art methods are based on Convolutional neural networks (CNN) structures. CNN, although successful, process only one local region at a time and can cause the detail loss of the data due to the existence of convolutional and down sampling operations. For the above problems, this paper designs a new goal-oriented Person ReID architecture, called Trans ReID. In this approach, the images are first converted into several patches, and strong baselines are built, which is advantageous against CNN-based method studies on several Person ReID datasets. To further enhance the robust features under Transformer for learning, two new modules are proposed. We use the movement and patch scrambling operations to rearrange the embedding of the patch for better identification performance and wider coverage, and we integrate and analyze some non-visual cues to reduce feature bias during camera view changes. The experimental results show our method achieves excellent performance on Person ReID datasets.

Keywords: TransReID · Target re-identification · Vision transformer

1 Introduction

Visual information accounts for about 80%–85% of the information obtained by the human perception system. Image and video related applications are increasingly prominent in the daily life of the people. Image processing is not only a challenging theoretical research direction in the science field, but also an important application technology in

the engineering field. Person ReID is a new technology emerging in the field of intelligent video analysis in recent years. It belongs to the category of image processing and analysis in complex video environment. It is the main task [1–3] in many monitoring and security applications, and has gained more and more attention to [4–8] in the field of computer vision.

Person ReID compares the pedestrian images taken by different non-overlapping cameras to further determine whether two pictures are the same person to establish the connection between different cameras in different areas, so as to realize pedestrian tracking. Person ReID technology has broad application scenarios, which is an important technical means for the public security to track criminal suspects, and also an important link in building intelligent security. To a large extent, Person ReID technology has changed the current situation of traditional manual observation of video surveillance frame by frame, realized intelligent video image analysis, and greatly reduced the cost of human labor and time. Therefore, the study of Person ReID has great social significance [18].

The study of Person ReID faces many challenges, such as low image resolution, visual angle change, attitude change, light change, and occlusion. For example, 1) The picture of the surveillance video is generally fuzzy, the resolution is also relatively low, as shown in Fig. 1(a), So the use of face recognition and other ways to recognition, only by using the human appearance information outside the head, different pedestrians may have the same body size and clothes, This brings great challenges to the accuracy of Person ReID; 2) Person ReID images are often taken from different video cameras, Due to the different shooting scenes and camera parameters, Person ReID work generally has problems such as light change and perspective change, as shown in Fig. 1(b) and (c). This results in large differences for the same pedestrian under different cameras, the physical characteristics of different pedestrians may be more similar than those of the same person; 3) Re-identified pedestrian images may be taken at different times, Pedestrian posture and clothing will be changed to varying degrees. In addition, the appearance characteristics of pedestrians will vary greatly under different light conditions, as shown in Fig. 1(c). In addition, the scene under the actual video surveillance is very complex, and many monitoring scenes have a large flow of people, complex scenes, and the picture is easy to block. See Fig. 1(d), when it is difficult to re-identify by gait and other characteristics. All the above situations bring great challenges to the study of Person ReID, so the current study is still far from at the practical application level.



(a)Low-resolution image (b) visual Angle change (c) light change (d) occlusion

Fig. 1. Difficulties and challenges of pedestrian re-identification

2 Related Works

Person ReID methods are mainly divided into: feature representation based method and metric based learning method. The feature representation-based method is mainly to learn a robust deep network extracting feature; the metric learning method mainly maps the pedestrian image to another space to make the distance of the same pedestrian less than the different pedestrian distance.

In this paper, through learning from CNN, two problems are not effectively solved, which have not been fully developed in the target ReID field. (1) For the target ReID, it is very important to apply the rich structural patterns to a global scale. However, due to the Gaussian distribution of the effective receiving field, the CNN technique focuses on identifying smaller regions. Recently, the attention module [9] has been introduced to study the remote dependency relationship of, but most modules are deeply embedded and cannot fundamentally solve the problem of CNN. So, while attention-based methods do have great advantages on a large scale, it is difficult to extract complex, diverse and differentiated parts see Fig. 2. In Fig. 2, (a) is the original image, (b) is a CNN-based method, (c) is a CNN attention learning method, and (d) is a transformation-based method. (1) Fine-grained characteristics are the key to obtain the ideal results. However, the down sampling operation reduces the resolution of spatial data that output a feature for mapping, and has a large impact on the resolution of similar objects. As shown in Fig. 3, some details of the backpack are lost in the CNN feature map, so it is difficult to distinguish between the two and it cannot be clearly identified.

Recently, Vision transformer (ViT) and Data-efficient image Transformers (DeiT) have shown that using traditional transformer for image recognition is comparable to the CNN features extraction. For two reasons, using the multi-head self-attention feature module, reducing convolution and down sampling operations, enables the transformer-based model to well solve the above problems in CNN-based ReID.

(2) Compared with the CNN model, multi-head-self-attention captures long-term dependencies and further drives the model to participate in different human parts (thighs, shoulders, and waist in Fig. 2) to capture global relevant information and some differentiated parts.

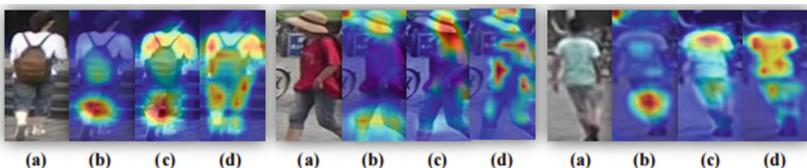


Fig. 2. Visual attention map [24]

(3) Without drop sampling operation, transformer can store more details. For example, we can observe that in different figures (marked with red squares in Fig. 3), we can easily distinguish them, and these advantages allow us to introduce tradition's transformer. (Left: CNN-based method, right: transformer-based method). In the red box, the input image is reduced to 1024×512 for better display as compared to the CNN-based method.



Fig. 3. Visualization of the output feature plots of 2 hard samples with similar appearance [24]

3 Methods of This Paper

The pedestrian ReID framework proposed in this paper comes from transformer-based image classification, but has several key improvements to capture robust features (Sect. 3.1). To further improve robust feature learning based on transformer, jigsaw patch module (JPM) and side information embedding (SIE) are carefully designed in Sects. 3.2, 3.3. These two modules are jointly trained in an end-to-end manner, as shown in Fig. 5.

3.1 Strong Baseline Based on Transformer

In this paper, we establish our own strong transformer based on the general strong pipeline of target ReID. It includes two different stages: feature extraction and supervised learning. Figure 4 gives an image $x \in R^{H \times W \times C}$. Its height, width, and number of channels are indicated by H, W, and C, respectively, and then split into N fixed-size patches $\{x_p^i | i = 1, 2, \dots, N\}$. An additional learnable embedding tag x_{cls} , expressed as before being added to the input data sequence. The output [cls] token is used as a global feature representing f. The integration of spatial data is achieved by adding learnable location embeddings. Then, the input sequence is entered into the transformer. Layers can be represented as:

$$z_0 = \left[x_{cls}; F(x_p^1); \dots; F(x_p^n) \right] + p \quad (1)$$

where Z_0 represents the image input sequence embedding, $p = R^{(N+1) \times D}$ is the position embedding. F is a linear projection that maps the patch to the D dimension. In addition, the i transformer layers were also used to learn the feature representation. The finite receptive field problem of the CNN-based methods is solved because all the transformer layers have their own one global acceptance field. The details were also preserved because of no downsampling.

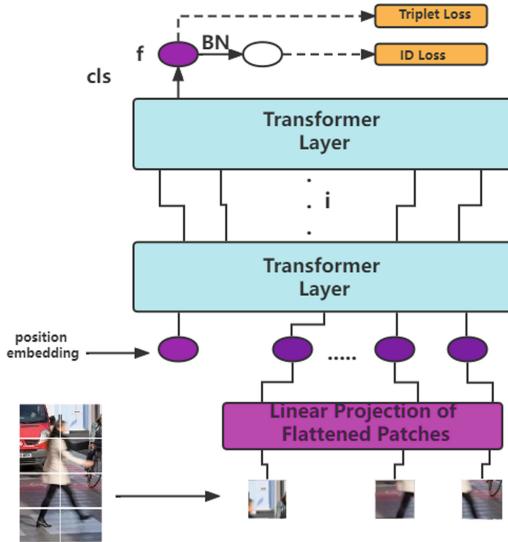


Fig. 4. Strong baseline framework based on transformer

Overlap patch. traditional transformer's model divides the images into non-overlapping patches, losing the local spatial proximity structure around the patches. In contrast, sliding windows are used to produce spots with overlapping pixels. The step size as S and the size of the patch as P (e.g.16), the shape of two adjacent patches is $(P-S) \times P$. An input image with an image resolution of $H \times W$ will also be split into N patches.

$$N = N_H \times N_W = \left\lfloor \frac{H + S - P}{S} \right\rfloor \times \left\lfloor \frac{W + S - P}{S} \right\rfloor \quad (2)$$

where, $\lfloor \cdot \rfloor$ is the maximum integer function, S is set to less than P . N_H and N_W indicate the number of patches formed at height and width, respectively. The smaller the S , the more patches the image is divided into. Simply put, the more patches, the better the performance, and the higher the computing costs position embedding. Because the image resolution of the ReID task is different from the original image resolution, it cannot be embedded in the ImageNet. Bilinear interpolation is used at any given input resolution, and the embedding technique for location information is equally learnable, just like ViT. The ID loss and ternary loss of the global features can be used to optimize the network. For a triad $\{a, p, n\}$, the soft-margin ternary L_T loss is as follows:

$$L_T = \log[1 + \exp(||f_a - f_p||_2^2 - ||f_a - f_n||_2^2)] \quad (3)$$

3.2 Jigsaw Patch Module

Although the robust baseline based on transformer presented in this paper can achieve good results on ReID, it utilizes the information from the entire image to study ReID. But because of occlusion and dislocation, usually only a small part of the object can be seen. Learning fine-grained local properties, such as band features, has gained a wide application in CNN to address the above problems.

Assuming that the hidden features are input to the last layer is represented as $Z_{l-1} = [z_{l-1}^0; z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$. An easy way to understand fine-grained local properties is to divide $[z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$, connect the shared markers z_{l-1}^0 in turn, and then input k feature sets into the common transformer layer, get k local features, and use $\{f_j^i | j = 1, 2, \dots, k\}$ and f_j^i as the output markers of the jth group. However, since each region only considers the local continuous patch embedding, the overall dependence of transformer can not be fully played. For the above problems, this paper proposes a puzzle Patch module (JPM) that is embedded in a patch to be patched and divided into several pieces, each containing several random pieces. Moreover, adding additional interference to the learning process can effectively improve the robustness of the target ReID model. By understanding the relevant studies of ShuffleNet, the patch embedding is shuffled by shift operations and patch shuffled operations. Sequence embedding Z_{l-1} is adjusted as follows:

Step 1: Shift-shift operation. The front block was moved to the back, the $[z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$ move m step changes to $[z_{l-1}^{m+1}, z_{l-1}^{m+2}, \dots, z_{l-1}^N, z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^m]$.

Step 2: Patch scrambling operation. In step 1, the moved block is further processed by a shuffled operation (group = k), where the implied feature changes to $[z_{l-1}^{x_1}, z_{l-1}^{x_2}, \dots, z_{l-1}^{x_N}]$, $x_i \in [1, N]$. The algorithm divides the obtained features into k groups, and JPM can divide these features into k parts $\{f_l^1, f_l^2, \dots, f_l^k\}$, so that each attribute can encode different components. Global and local features were trained using classification loss L_{ID} , L_T respectively, and gives the definition of overall loss:

$$L = L_{ID}(f_g) + L_T(f_g) + \frac{1}{k} \sum_{j=1}^k (L_{ID}(f_l^j) + L_T(f_l^j)) \quad (4)$$

In the prediction stage, this paper splicing the convolutional and local features to obtain the final feature expression $[f_g, f_l^1, f_l^2, \dots, f_l^k]$. If only used f_g , the operation can be greatly reduced, but the performance will also be reduced.

3.3 Auxiliary Information Embedding

When the fine-grained features are expressed, these features will still become indistinguishable because of the camera or perspective changes. That is, the trained model cannot identify the same target from multiple angles. In order to solve the above problems, in the CNN-based scheme design, it is often necessary to adjust the structure of the network, or adjust the loss function, to maximize the use of non-visual image information (e.g. camera ID, angles, etc.). This paper introduces an auxiliary information embedding

(SIE) technology, which beds non-visual information in an image such as camera and angle into the image to learn the invariance of the image.

Transformer is very good at integrating this boundary information, and the auxiliary information embeddings shown in Fig. 5 encode embedded non-visual information such as cameras or angles. The method is input into the encoder of a common transformer, together with the embedding of the patch, the embedding of the location. The last layer consists of two layers of non-affecting transformer. One is to encode the overall features with standard methods. The other is a puzzle patch module (JPM) that can clear and regroup all the patches. To better understand the local function, all groups were placed into a common transformer layer. Both global and local features can easily cause the loss of ReID, and the method inserts SIE together with patch embedding and location embedding into the transformer encoding device to encode these boundary information, so Transformer is useful when handling ReID tasks. Specifically, if the camera ID of an image is C, the corresponding camera embedding can be represented by $S(C)$, in different locations, the camera embedding changes with the pixel block, and the camera embedding is the same for all blocks. Furthermore, the method can encode all blocks when the angle label V is already known.

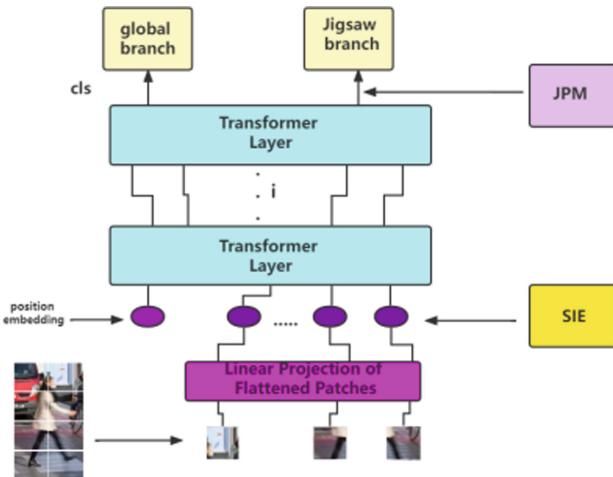


Fig. 5. The TransReID's framework

The next step is to think about how to integrate two different pieces of information. The simplest way is to add it up directly, $S(C) + S(V)$, But this approach may cause an information offset. This paper presents a new coding scheme: $S(C, V)$. In other words, There are C_N camera ID and V_N angle tags, produces $C_N \times V_N$ values. Then, the ith data block is defined as follows:

$$Z'_0 = Z_0 + \lambda S_{(C,V)}[r * N_V + q] \quad (5)$$

4 Experiments

In this paper, we test the performance of TransReID through several experiments and derive the values of Rank-n of TransReID as well as mAP on several common datasets. Comparast tests were also performed on the DukeMTMC-reID dataset, where these parameters were obtained to obtain the effect of JPM and SIE modules on improving performance rates. Comparisons with other complex networks are also discussed to illustrate the strengths of our TransReID.

4.1 Dataset and Composition

We evaluate the proposed method on the Market1501, DukeMTMC-reID, MSMT17, VeRi-776, dataset. These datasets all provide a camera ID for each image. These details are summarized in Table 1.

Table 1. Statistics of the datasets used in this article

Data set	Object	#ID	#image	#cam	#view
Market1501	Person	1, 501	32, 668	6	-
MSMT17	Person	4, 101	126, 441	15	-
DukeMTMC-reID	Person	1, 404	36, 441	8	-

4.2 Set up the Parameters and the Operating Environment

The specific experimental operating environment is shown in Table 2:

Table 2. Model training environment

CPU: i7-7700HQ	OS: Window10
GPU:Nvidia Geforce 3080	CUDA: 11.0
Memory: 32G	yacs 0.16
Python: 3.6.5	timm: 0.5.4
opencv-python: 4.1.0.25	Pytorch: 1.7
Torchvision: 0.8.0	Geforce Graphics Driver: 512.59

Important parameter settings during training are shown in Table 3 (take Market1501 as an example):

Table 3. Network training parameters

Class	Set value
STRIDE_SIZE	[12, 12]
METRIC_LOSS_TYPE	Triplet
MAX_EPOCHS	120
IMS_PER_BATCH	256

All pedestrian images were adjusted to 256×128 . Training images were enhanced by random horizontal flipping, padding, random cropping, and random erasing. The batch size was set to 64 with 4 images per ID. The SGD optimizer was used with a momentum of 0.9 and 1e-4 for weight decay. The learning rate was initialized to 0.008, and the cosine learning rate decayed. Set $m = 5$ and $k = 4$ for the human ReID dataset. The initial weights of ViT were pre-trained on ImageNet-21K and then fine-tuned on ImageNet-1K, while DeiT was trained only on ImageNet-1K. Following the conventions of the ReID community, in this paper all methods were evaluated using the CMC curves and mAP.

4.3 Experiment Comparison

1) Results from the transform-based baseline

In Table 4, the CNN-based and transformer-based backbone networks are compared (take MSMT17 as an example), and several different backbone networks are selected to show their balance between computing and performance. The predicted time consumption for each backbone was also included for a comprehensive comparison.

It can be seen that the Deit-S/16 has slightly better performance and speed than the ResNet50. DeiT-B/16 and ViT-B/16 achieve similar performance on ResNeSt50 when shorter than ResNeSt50 (1.79x vs 1.86 x), while the baseline performance improves and the prediction time decreases the sliding window S.ViT-B/16s = 12 is faster than ResNeSt200(2.81x vs 3.12x) and slightly better than ResNeSt200 on the ReID benchmark.

Table 4. Comparison between the different Backbones

Backbone	Inference time	MSMT17 mAP R1
ResNet50	1x	51.3 75.3
ResNet101	1.48x	53.8 77.0
ResNet152	1.96x	55.6 78.4
ResNeSt50	1.86x	61.2 82.0
ResNeSt200	3.12x	63.5 83.5
DeiT-S/16	0.97x	55.2 76.3
DeiT-B/16	1.79x	61.4 81.9
ViT-B/16	1.79x	61.0 81.8
ViT-B/16s = 14	2.14x	63.7 82.7
ViT-B/16s = 12	2.81x	64.4 83.5

Table 5. Effectiveness study of the JPM module

Backbone	#groups	MSMT17 mAP R1
Baseline	-	61.0 81.8
JPM	1	62.9 82.5
JPM	2	62.8 82.1
JPM	4	63.6 82.5
JPM w/o rearrange	4	63.1 82.4
JPM w/o local	4	63.5 82.5

2) Compare the effectiveness of the JPM module.

Table 5 In MSMT17, JPM brings a 2.6% mAP improvement compared to baseline. Also increasing the number of groups k improves performance while slightly increasing the inference time. In experiments, the group number k = 4 is a trade-off between speed and performance. Comparing JPM and JPM w/o rearrange (without rearrangement sequence), we can observe that the shift and shuffle operations obtain 0.5% mAP improvement in MSMT17, which helps the model to learn more discriminant features. It is also observed that evaluating global features without connecting locally, the performance is almost comparable to the version of the full feature if only global features fg is used in the inference phase, so it is recommended to use only global features as an efficient variant with lower storage and computational cost in the inference phase

Fig. 6. The attention visualization shows ((a) input image, (b) baseline, (c) JPM w/o rearrange, (d) JPM), JPM with rearrange operation can help the model learn more global context information and more discriminant parts, which makes the model more robust to perturbations.

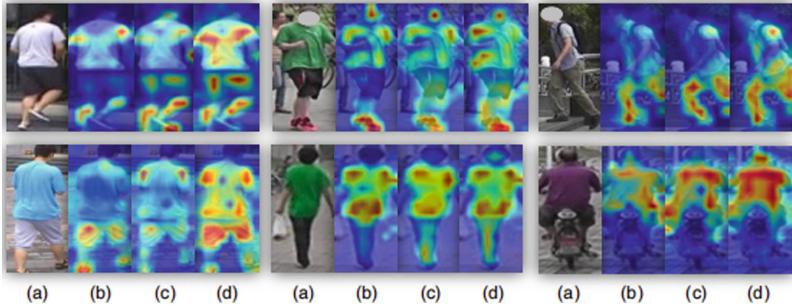


Fig. 6. Visual attention map [18]

3) Compare the effectiveness of the SIE.

In Table 6, the effectiveness of the SIE is evaluated. The MSMT17 does not provide viewpoint annotation, specifically using the VeRi776 dataset, the VeRi-776 not only has the camera ID for each image, but also labels eight different viewpoints based on the vehicle orientation. Thus, the results are displayed by various combinations of SIE encoding the camera ID and/or angle labels. It can be seen that the addition of camera ID mAP by 0.5%, the addition of angle tag mAP by 0.3%, and the 1.4% mAP.

Table 6. Study on the effectiveness of SIE

Method	Camera viewpoint	VeRi-776 mAP R1
Baseline	✗ ✗	78.2 96.5
SC [r]	✓ ✗	78.7 97.1
SV [q]	✗ ✓	78.5 96.9
S (C, V)	✓ ✓	79.6 96.9

The effect of the weights of the SIE modules on the performance is analyzed in Fig. 7. When 0, the baseline reached 61.0% mAP on the MSMT17. With the increase, the mAP increases to 63.0% (=2.0 for MSMT17), which means that the SIE module now has the advantage of learning invariant features. However, after more than 2, performance decreases because the weight of feature and location embedding decreases.

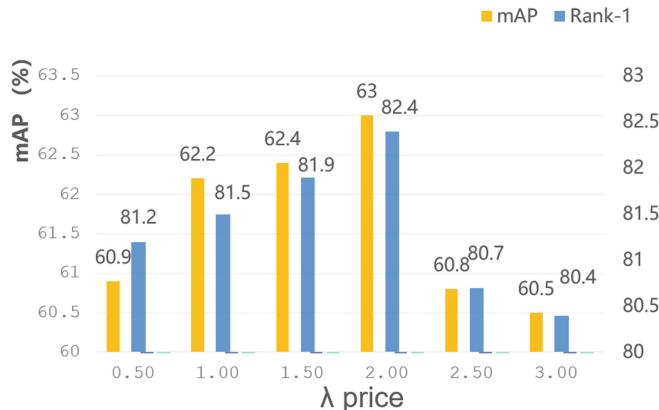


Fig. 7. Effect of the change on the MSMT17

4) Compare the effectiveness of the TransReID.

Table 7 fully illustrates the advantages of adding the JPM module and the SIE modules. For the baselines, JPM and SIE improved the performance of DukeMTMC-reID 0.6%mAP, MSMT172.6% and 1.4%, and Market15011.5% and 1.4%.

Table 7. Effectiveness of TransReID

Method	JPM SIE	DukeMTMC-reID mAP R1	MSMT17 mAP R1	Market1501 mAP R1
Baseline	✗ ✗	79.7 88.9	61.0 81.8	86.8 94.7
	✓ ✗	80.3 89.6	63.6 82.5	88.3 95.0
	✗ ✓	80.3 89.5	62.4 81.9	88.2 95.0
TransReID	✓ ✓	81.2 90.0	67.8 85.3	88.9 95.0

After using these two functional modules together, the mAP of 81.2%, 67.8 and 88.9 was TransReID on Market1501 in DukeMTMC-reID and MSMT17. It can be seen that TransReID has the worst performance on MSMT17, as shown in Fig. 8.

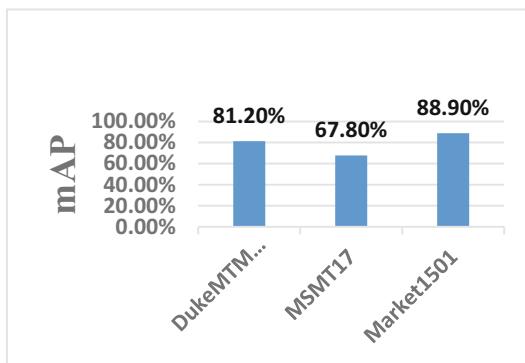


Fig. 8. The mAP on the pedestrian data set

5 Conclusion

This paper designs transformer for target ReID tasks and proposes two new modules: puzzle patch module (JPM) and Auxiliary Information Embedding (SIE). Comparative experiments on the DukeMTMC-reID dataset can fully show the effective enhancement of accuracy by the JPM and SIE modules. At the same time, TransReID has achieved good performance on multiple datasets, and the results indicate the great application value of this transformer in the ReID task. The disadvantage is that by comparing the accuracy of TransReID on DukeMTMC-reID, MSMT17, Market1501, we can see in the MSMT17 accuracy of the lowest, the analysis because the MSMT17 picture environment is complex, such as different weather, different light, these factors affect the final accuracy, the next step of this paper will be how to deal with the influence of these factors.

Acknowledgment. This work was supported by the open fund of Information Materials and Intelligent Sensing Laboratory of Anhui Province with Grant No. IMIS202114, by the second batch of industry university cooperation collaborative education project of the Ministry of education in 2021 with No. 202102326028, by key scientific research projects of Suzhou University in 2021 with No. 2021yzd01 and 2019yzd05.

References

1. Li, Y., Wu, Z., Karaman, S., et al.: Real-world re-identification in an airport cameranetwork. In: International Conference on Distributed Smart Cameras, Venice, Italy, p. 35 (2014)
2. Gong, S., Cristani, M., Yan, S., et al.: Person Re-identification. Springer, London (2014)
3. Camps, O., Gou, M., Hebble, T., et al.: From the lab to the real world: Re-identification in an airport camera network. *IEEE Trans. Circuits Syst. Video Technol.* **2016**(99), 540–553 (2016)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision, Marseille, France, pp. 262–275 (2008)
5. Prosser, B., Zheng, W.S., Gong, S., et al.: Person re-identification by support vector ranking. The British Machine Vision Conference. Aberystwyth, British, pp. 1–21 (2010)

6. Jurie, F., Mignon, A.: PCCA:a new approach for distance learning from sparse pairwise constraints. In: IEEE Conference on Computer Vision and Pattern Recognition.IEEE Computer Society, pp. 2666–2672 (2012)
7. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. Oregon, USA, pp. 3586–3593 (2013)
8. Zheng, W.S., Li, X., Xiang, T.: Partial person re-identifification. In: IEEE International Conference on Computer Vision.Santiago, Chile, pp. 4678–4686 (2015)
9. Chen, X., et al.: Salience-guided cascaded suppression network for personre-identification. In: CVPR, June 2020. 1,8



Modified Lightweight U-Net with Attention Mechanism for Weld Defect Detection

Lei Huang¹, Shanwen Zhang^{2(✉)}, Liang Li¹, Xiulin Han³, Rujiang Li³,
Hongbo Zhang¹, and Shaoqing Sun³

¹ Tubular Goods Research Institute of CNPC, Xi'an 710077, China

² College of Electronic Information, XiJing University, Xi'an 710123, China
huanglei002@c npc .com .cn

³ North China Petroleum Steel Pipe Co., Ltd., Qingxian 062653, Hebei, China

Abstract. Welding is an important joining technology but the defects in welds wreck the quality of the product evidently. Weld defect detection is always an important and challenging research due to the various defects with complex background. A modified lightweight U-Net with attention mechanism model (LWAU-Net) is constructed for weld defect detection. In the model, the multiple convolutional and pooling kernels with different sizes are utilized to learn the multi-scale discriminant features, and the attention mechanism is used to capture to adaptively select multi-scale features for classification. Compared with the standard convolution neural networks (CNN), LWAU-Net is integrated to learn the multi-scale features for weld defect detection, especially small defects. Experiment results on the weld defect image dataset show that the proposed method outperforms the state-of-the-art method on the same dataset and the obtained multi-size defect edge is clearer.

Keywords: Weld defect detection · U-Net · Attention mechanism · Lightweight U-Net with attention mechanism (LWAU-Net)

1 Introduction

Weld defect detection is always an important research in computer vision and in the manufacturing process of large-scale equipment. As an important process in the production of construction machinery, welding is a complex process involving arc physics, metal phase transformation, heat transfer and mechanics, which is directly related to the quality and safety of construction machinery. Due to highly concentrated instantaneous heat input during welding process, considerable welding residual stress and deformation are easy to occur after welding [1, 2]. The weld defects are divided into surface defects and internal defects. The surface defects include welding tumor, spatter, sag, surface crack and so on. Surface defects tend to produce stress concentration points after the welding load, or reduce the effective cross-sectional area of the weld and reduce the strength of the weld. Besides, the welding process is generally clearly defined, and visual inspection of surface defects can be used. Internal defects have porosity, slag inclusion, cracks,

incomplete fusion, etc., these problems will affect the quality of the welding, reduce the mechanical properties of welding, caused a lot of waste and defective products, causes a lot of trouble to the life and production safety hidden trouble, internal mainly detected by non-destructive testing to detect weld defect detection system at the time of welding defects of image, the influence of various interference factors due to the outside world [3, 4]. The defect feature of the obtained defect images is not obvious and their edge is not clear. Affect the normal judgment and recognition of welding defects. Defects in the welding process have the possibility of expansion and even catastrophic consequences. Therefore, detecting the defects of welding structure in advance can ensure the safe operation of equipment [5, 6].

Welding is a very complicated process with uncertainty and multi-variable influence. The welding quality cannot be effectively guaranteed due to the instability of welding personnel's actual operation and the influence of complex construction environment. Therefore, effective inspection of welding quality, that is, accurate detection of weld defects, is of great significance to ensure the quality of weld and industrial safety production. Many weld defect detection methods based on image processing have been proposed. In order to facilitate the defect identification of weld defect images, image processing is generally used to suppress noise and improve image quality. Image analysis is to segment the image. Edge detection technology is the basis of image analysis and processing, such as image segmentation, image enhancement, image restoration, pattern recognition, image compression, etc. It is the first step of all image analysis methods based on edge segmentation. There are many edge detection methods based on gray image [7], such as the commonly used Roberts operator, Sobel operator, Prewitt operator, Laplacian operator and Canny operator, etc. Although the defects detected by these methods are feasible, it is easy to lose small defect information. Many scholars use the edge detection method of noisy color image based on morphology, which does not blur the edge details of defective image and does not affect the extraction of image features in the process of image filtering [8, 9].

To improve the accuracy of automatic defect classification, Mu et al. [10] developed a weld defect classification algorithm based on principal component analysis (PCA) and support vector machine (SVM). Later, Mu et al. [11] proposed a weld defect classification approach based on dual-parameters optimization of the principal component analysis (PCA) and the linear discriminant analysis (LDA). The original defect images are transformed to eigen-defects by PCA and LDA is used to classify eigen-defects. The optimal parameters of PCA and LDA are given when the PCA-LDA model gets the maximum value of classification accuracy. Murugan et al. [12] studied on the effect of weld defect on the fatigue behavior of welded structures and aimed at prediction of fatigue behaviour of symmetric structures like pressure vessels in the presence of common welding defects such as lack of fusion (LOF), lack of penetration(LOP)and porosity. Among required weld qualities, the weld joint penetration is often the most critical one as an incomplete penetration causes explosion under high temperature/pressure and an excessive penetration/heat input affects the flow of fluids and degrades materials properties. Cheng et al. [13] proposed formed a composite image from the image taken from the initial pool, reflecting prior condition and from real-time developing pool such that this single composite image reflecting the measurable phenomena is only determined by

the development of the weld penetration. Yazid et al. [14], presented a simple yet robust algorithm for texture identification using Dimensional Discrete Fourier Transform and Dynamic Time Warping with illumination variations. Wei et al. [15] aimed to provide the technical basis for the automatic tracking of weld joints from the face panel side, required for the high-reliability manufacturing of curved sandwich structures, proposed a weld position detection method based on backscattered X-ray, and carried out several experiments on a 6061 aluminum alloy specimen with a thickness of 3 mm. The experimental results demonstrate that the maximum absolute error of the detection was 0.340 mm, which is sufficiently accurate for locating the position of the joint. Zhang et al. [16] tried to analyze the defect signal characteristics and discuss the extraction methods by compiling the image preprocessing procedure, columns, gray-scale extraction procedure and defect alarm procedures, and draw the conclusion that column gray method can be used as a dynamic test basis of automatic alarm decision.

Traditional methods generally include several serial steps, such as image preprocessing, region segmentation, feature extraction, and type recognition. The results of each step have significant impact on the accuracy of the final defect identification. Convolutional neural network (CNN) has become dominant in various computer vision tasks [17]. It is mainly composed of these types of layers: input layer, convolutional layer, ReLU layer, pooling layer, and fully connected layer (the fully connected layer is the same as the conventional neural network). It is suitable for this problem to classify the variation of each weld defect pattern. Khumaidi et al. [18] proposed a welding defect classification method using CNN and Gaussian kernel, where CNN consist of two stages: extraction image using image convolution and image classification using neural network. Zhang et al. [19] explored a possible solution for weld defect detection and proposed an image-based approach using small X-ray image data sets. They trained two deep CNNs on the augmented image sets using feature-extraction based transfer learning techniques. The two trained CNNs are combined to classify defects through a multi-model ensemble framework, aiming at lower false detection rate. Zhang et al. [20] constructed a deep learning network for X-ray weld defect detection, and analyzed the size and number of layers of CNN template, and the influence of different activation functions. Liu et al. [21] proposed a signs recognition framework by CNN for weld images and presented a spatial and channel enhancement (SCE) module to enhance the practicability of the framework. Jiang et al. [22] proposed a CNN-based weld defect recognition method, which includes an improved pooling strategy and an enhanced feature selection method. According to the characteristics of the weld defect image, an improved pooling strategy that considers the distribution of the pooling region and feature map is introduced.

Weld defect detection is still an important and challenging task due to the various defects, as shown in Fig. 1. Inspired by the above methods based on CNN and U-Net and their modified models, this paper constructs a lightweight U-Net with attention (LWAU-Net) model for weld defect detection. The experimental results on a weld defect detection image dataset demonstrate the validity and generality of the proposed detection algorithm.

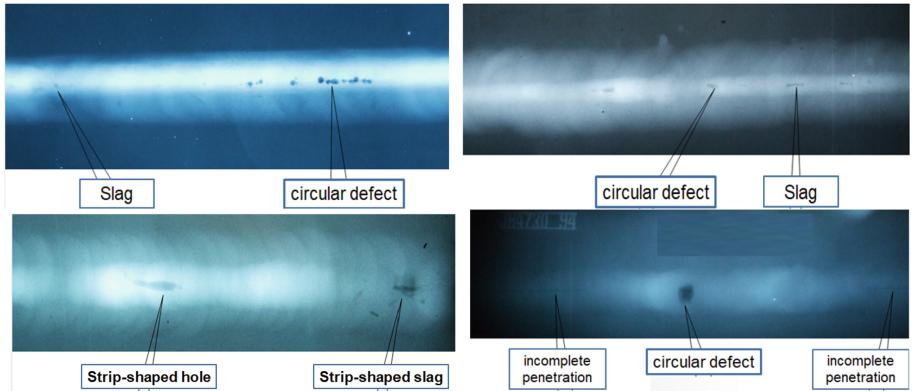


Fig. 1. Weld defect examples

2 U-Net

U-Net is an encoder-decoder symmetric network composed of convolution, down-sampling, up-sampling and splicing operations. A typical U-Net structure is shown in Fig. 2.

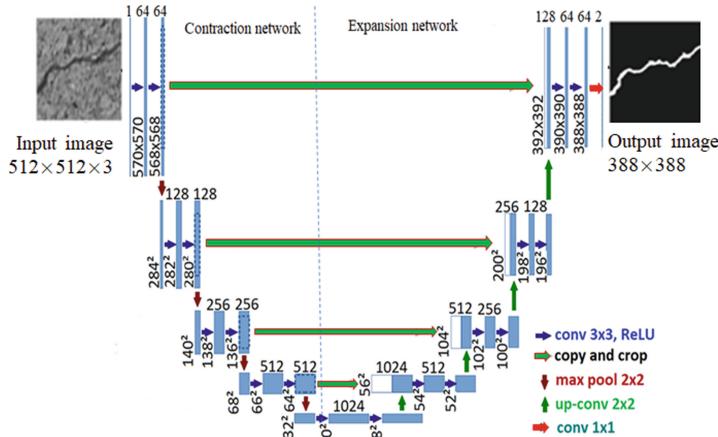


Fig. 2. U-Net structure

U-Net network is composed of four blocks, each containing two 3×3 convolution, ReLU and one downsampling operation. The purpose of using 3×3 convolution kernel is to minimize the complexity of neural network and maintain the segmentation accuracy. The left half of the network is the contraction path, and the right half is the expansion path. The U-Net network extension path gradually repaired the details of the image, accurately located the location of the lesion, and restored the image to the image size. The extension path also contains four blocks, each of which contains two valid convolution, ReLU and one upsampling operation. The upsampling operation is to restore and decode the abstract

features of the image obtained from the downsampling operation to the original size. After each upsampling operation, the size of the feature map is doubled and the number of channels is halved. A jump connection is added between the contraction path and the expansion path for pixel location. Different from the summation of FCN, U-Net uses concatenation, the feature map of the shrink path of the same layer is cut to the same size as the extended path, and then the splicing operation is helpful to restore the information loss in the process of lower sampling.

3 Lightweight U-Net with Attention (LWAU-Net)

Due to the variety of shapes and sizes of RSI, U-Net inspection defect alone cannot meet the requirements for accuracy, speed, etc. Aiming at the problem of low contrast between defect target area with irregular shape and size and difficult segmentation, the symmetrical encoder-decoder structure of U-Net network has achieved good results in the field of image segmentation, and many image segmentation networks have been improved based on it. The attention mechanism is introduced into U-Net to build an improved lightweight U-Net model (LWAU-Net), whose structure is shown in Fig. 3A. The skip connection in U-Net is replaced by Respath as shown in Fig. 3B. The attention mechanism is shown in Fig. 3C.

In U-Net, the discriminant ability of high-level feature maps is strong, but the target is vague. The target localization of low-level feature map is more accurate, but the classification ability is poor, and there are more virtual target information and scene content. To integrate the advantages of high-level feature maps and low-level feature maps, an attention mechanism is added to U-Net to generate rough attention maps. In U-Net, attention can generally be divided into channel attention module and spatial attention module, as shown in Fig. 3C. Channel attention can select channels to represent different feature information of images, and spatial attention is used to select the regions to be noticed in the image feature map.

LWAU-Net is an improved U-Net model, and the improvement includes data enhancement, convolution operation, down-sampling operation, up-sampling operation, and model optimization strategy and jump connection.

The encoder of LWAU-Net adopts multi-scale convolution model Inception V1, Inception has four basic components: 1×1 convolution, 3×3 convolution, 5×5 convolution and 3×3 maximum pooling. Different convolutional kernels of different sizes are used to achieve different scales of perception, and finally, better image representation can be obtained through fusion.

The convolutional operation of LWAU-Net extracts useful features from the input image. Activation function and normalization are used to accelerate network convergence and solve the problem of gradient disappearance after the convolutional operation of U-Net. The improvement of convolution operation mainly includes the improvement of convolution block and convolution filling method. Convolutional block improvements include asymmetric convolution, dilated convolution and the addition of Inception module. Asymmetric convolution is the dismantling of an N by n convolution into an N by 1 convolution, followed by a 1 by n convolution in series. For example, 3×3 standard convolution can be decomposed into 1×3 and 3×1 single convolution operation, which can obtain the same effect as 3×3 convolution, but reduce the amount of computation

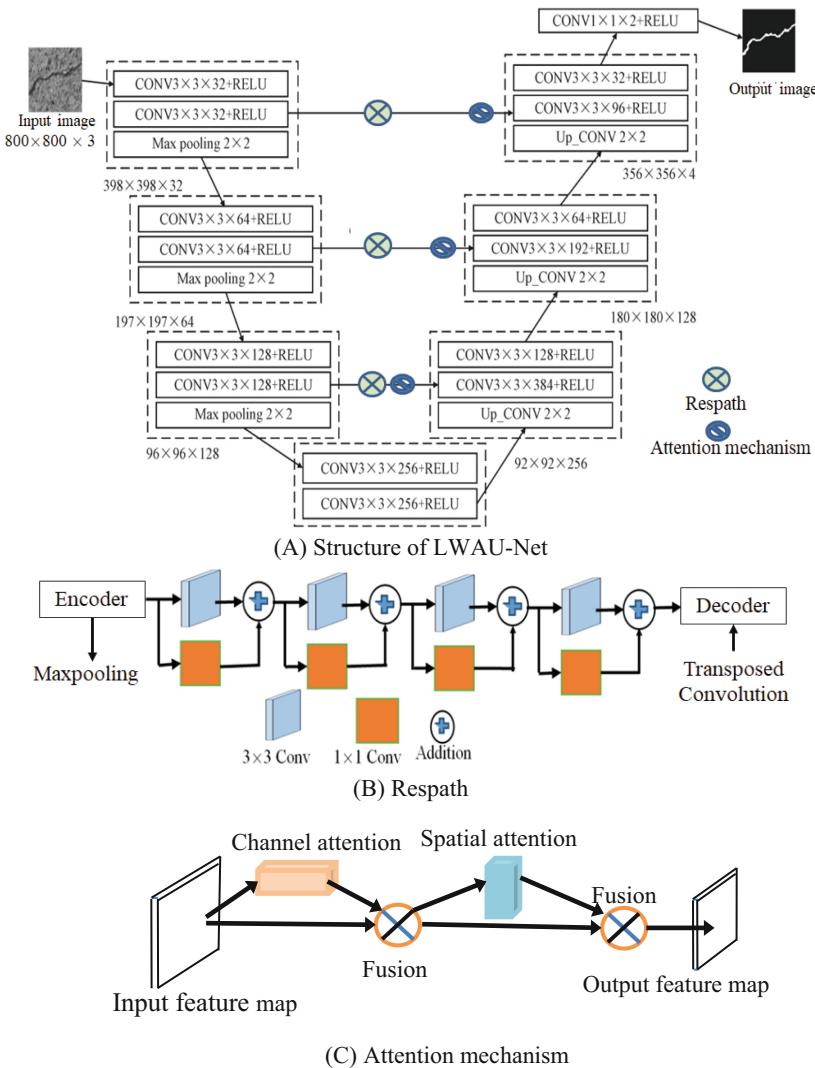


Fig. 3. Lightweight U-Net with attention (LWAU-Net)

and network parameters. Expansion convolution is based on the original convolution kernel, sampling is carried out by adjusting the expansion rate, and the receptive field is increased without increasing the number of calculated parameters, so that each convolution output contains a wide range of feature information. There are two disadvantages of dilated convolution. (1) After multiple superposition of dilated convolution, all pixels in the image cannot be applied due to discontinuous kernel, resulting in discontinuous information. (2) Using expansion convolution to segment relatively large targets in the image can improve speed and accuracy, but for relatively small targets in the image, the performance of empty convolution will be reduced.

The upsampling operation of LWAU-Net aims to restore the feature graph after feature extraction by convolution operation to the original size, which mainly includes transpose convolution, nearest neighbor interpolation, bilinear interpolation, trilinear interpolation and sub-pixel convolution. The advantage is that the calculation is convenient and simple, but the disadvantage is that the image is easy to appear jagged. Bilinear interpolation makes full use of the four real pixel values around the virtual points in the source graph to perform linear interpolation in two directions. First, linear interpolation is performed for pixel points A and B, C and D in the x direction to obtain M and N respectively. Then the linear interpolation of m and n of pixel points in the y direction is carried out to obtain the pixel interpolation of p point. It can be seen that linear interpolation in either direction gives the same result.

Maximum pooling is used in the downsampling process of LWAU-Net to further filter the features after convolution and reduce the parameters and computation, but it will cause the loss of spatial information. The improvement of pooling layer mainly includes maximum pooling, average pooling, random pooling, span convolution, dilated convolution, spatial pyramid pooling and Inception module.

The jump connection of LWAU-Net cuts and adds the feature graph after the convolution of the coding path to the decoding path, and introduces the attention mechanism to realize the positioning of the defect pixels. The methods of jump connection include attention mechanism block, feature reuse and attention mechanism block (FRAM), deconvolution + activation function, annotation information obtained from Siam network and new jump connection method.

4 Experiments and Analysis

To validate LWAU-Net based weld defect detection method, a lot of experiments are conducted on a weld defect image dataset, which was provided by CNPC Tubular Goods Research Institute, including crack, porosity, lack of penetration and lack of fusion, respectively, as shown in Fig. 4. Before training, these labeled images need to be processed into black and white binary images, namely mask images. Among them, the white pixels in the mask image correspond to the neural region in the ultrasound image, and the black pixels correspond to the non-neural background region in the ultrasound image.

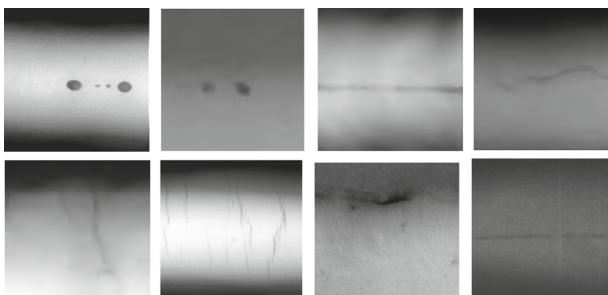


Fig. 4. Weld defect image examples

The convolutional processing image examples of LWAU-Net are shown in Fig. 5.

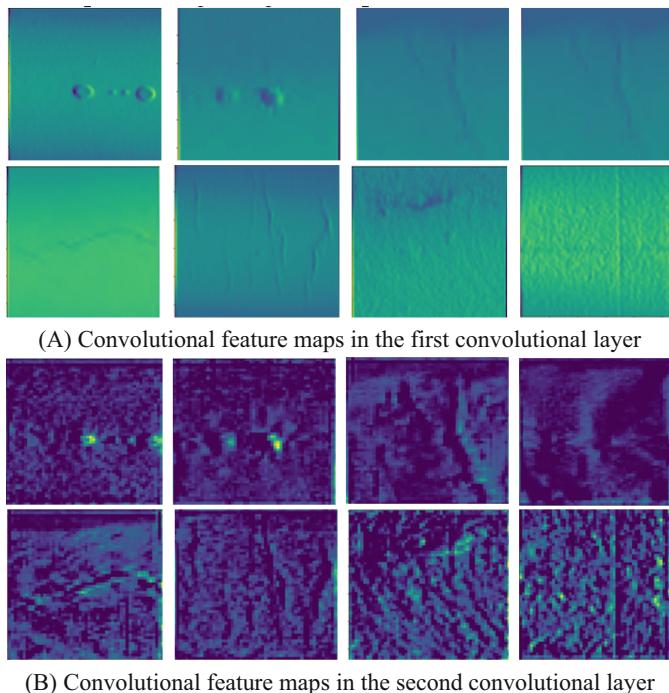


Fig. 5. The convolutional processing image examples of LWAU-Net

The detection results are shown in Fig. 6.

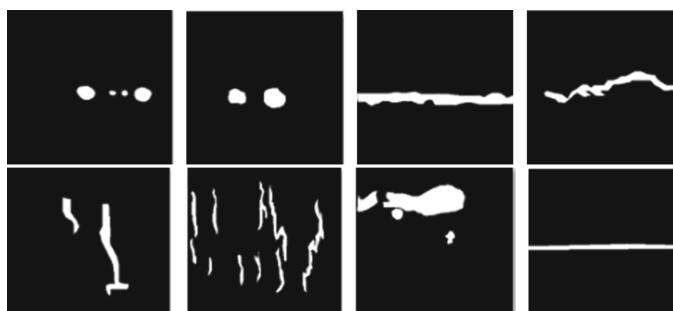


Fig. 6. The detection results

500 weld images are used in the experiment, 100 images per kind of defect, including 100 images without defects. Four-fold cross validation strategy is adopted. The trained network is used for defect detection, and the final output result of the network is the

probability of defect, as shown in Table 1. It can be seen that all defect-free images have been detected successfully with an accuracy of 100%. The accuracy rate of weld defect detection is 97%.

Table 1. Weld defect detected results

Weld defect	Number of weld defect samples	Number of defect errors	Accuracy /%
No defect	25	0	100
Crack	25	0	100
Porosity	25	1	96
Lack of penetration	25	1	96
Lack of fusion	25	1	96

As shown in Fig. 6 and Table 1, it is found that LWAU-Net has a great advantage in small weld defect detection.

5 Conclusions

Weld defect detection is always an important but challenging task. The standard CNN, FCN and U-Net can extract rich semantic features, but they may loss the bottom-level location information. The features of small defects may also be submerged by redundant top-level features, resulting in poor detection. To improve the weld defect detection algorithm, a modified lightweight U-Net with mechanism model namely LWAU-Net is constructed for weld defect detection. Comparison results show that LWAU-Net is effective and feasible for weld defect detection. How to overcome the following two challenges is the future work: (1) Inadequate generalization ability of segmentation model leads to the difficulty of the same network to have the same performance in multiple medical segmentation tasks. (2) U-Net has achieved a good segmentation effect on remote sensing image weld defect detection, but it faces the same problem of low interpretability as other deep learning methods.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Nos. 62172338 and 62072378).

References

1. Gang, T., Takahashi, Y., Wu, L.: Intelligent pattern recognition and diagnosis of ultrasonic inspection of welding defects based on neural network and information fusion. *Sci. Technol. Weld. Joining* **7**(5), 314–320 (2002)

2. Dinhm, M., Gu, F.: Autonomous weld seam identification and localization using eye-in-hand stereo vision for robotic arc welding. *Robot. Comput.-Integr. Manuf.* **29**(5), 288–301 (2013)
3. Zhang, L., Ye, Q., Yang, W., et al.: Weld Line Detection and Tracking via Spatial-Temporal Cascaded Hidden Markov Models and Cross Structured Light. *IEEE Trans. Instr. Measur.* **63**(4), 742–753 (2014)
4. He, Y.: Weld seam profile detection and feature point extraction for multi-pass route planning based on visual attention model. *Robot. Comput. Integr. Manuf.* **37**, 251–261 (2015)
5. Leemans, V., Destain, M.F.: Line cluster detection using a variant of the Hough transform for culture row localization. *Image Vis. Comput.* **24**(5), 541–550 (2016)
6. Shao, W.J., Huang, Y., Zhang, Y.: A novel weld seam detection method for space weld seam of narrow butt joint in laser welding. *Optics Laser Technol.* **99**, 39–51 (2018)
7. Akram, J., Kalvala, P.R., Chalavadi, P., Misra, M.: Dissimilar metal weld joints of P91/Ni alloy: microstructural characterization of HAZ of P91 and stress analysis at the weld interfaces. *J. Mater. Eng. Perform.* **27**(8), 4115–4128 (2018). <https://doi.org/10.1007/s11665-018-3502-8>
8. Kulkarni, A., Dwivedi, D.K., Vasudevan, M.: Study of mechanism, microstructure and mechanical properties of activated flux TIG welded P91 Steel-P22 steel dissimilar metal joint. *Mater. Sci. Eng. A.* **731**, 309–323 (2018)
9. Sirohi, S., Kumar, S., Bhanu, V., et al.: Study on the Variation in Mechanical Properties along the Dissimilar Weldments of P22 and P91 Steel. *J. Mater. Eng. Perform.* **31**, 2281–2296 (2022)
10. Mu, W., Gao, J., Jiang, H., et al.: Automatic classification approach to weld defects based on PCA and SVM. *Insight Non Destructive Testing & Condition Monitoring* **55**(10), 535–539 (2013)
11. Mu, W., Liu, G., Peng, L., et al.: A novel classification approach of weld defects based on dual-parameters optimization of PCA and LDA. In: International Conference on Advances in Mechanical Engineering & Industrial Informatics, pp. 1425–1429 (2015)
12. Murugan, R., Venugobal, P.R., Ramaswami, T.P., et al.: Studies on the effect of weld defect on the fatigue behavior of welded structures. *China Weld.* **27**(1), 53–59 (2018)
13. Cheng, Y., Wang, Q., Jiao, W., et al.: Detecting dynamic development of weld pool using machine learning from innovative composite images for adaptive welding. *J. Manuf. Process.* **56**, 908–915 (2020)
14. Yazid, H., Arof, H., Yazid, H., et al.: Weld detect identification using texture features and dynamic time warping. *Appl. Mech. Mater.* **752–753**, 1045–1050 (2015)
15. Wei, A., Chang, B., Xue, B., et al.: Research on the weld position detection method for sandwich structures from face-panel side based on backscattered X-ray. *Sensors* **19**(14), 3198 (2019)
16. Zhang, P.L., Zhao, Z.Q., Wang, Y.P.: X-Ray testing of weld defect of automatic recognition and alarm technology research. *Appl. Mech. Mater.* **472**, 495–502 (2014)
17. Ren, J., Wang, Y.: Overview of object detection algorithms using convolutional neural networks. *J. Comput. Commun.* **10**, 115–132 (2022)
18. Khumaidi, A., Yuniarno, E.M., Purnomo, M.H.: Welding defect classification based on convolution neural network (CNN) and Gaussian kernel. In: International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 261–265 (2017)
19. Zhang, H., Chen, Z., Zhang, C., et al.: Weld defect detection based on deep learning method. In: IEEE 15th International Conference on Automation Science and Engineering (CASE), pp. 1574–1579 (2019)
20. Zhang, L.F., Gao, W.X., Wang, Z., et al.: Research on weld defect identification with X-ray based on convolutional neural network. *J. Phys: Conf. Ser.* **1894**(1), 012071 (2021)
21. Liu, M., Xie, J., Hao, J., et al.: A lightweight and accurate recognition framework for signs of X-ray weld images. *Comput Ind* **135**, 103559 (2021)

22. Jiang, H., Hu, Q., Zhi, Z., et al.: Convolution neural network model with improved pooling strategy and feature selection for weld defect recognition. *Welding in the World, Le Soudage Dans Le Monde*, 65(4), pp. 731–744 (2020)



Handwritten Chemical Equations Recognition Based on Lightweight Networks

Xiao-Feng Wang¹, Zhi-Huang He¹, Zhi-Ze Wu², Yun-Sheng Wei³, Kai Wang¹, and Le Zou^{1(✉)}

¹ Anhui Provincial Engineering Laboratory of Big Data Technology Application for Urban Infrastructure, School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, Anhui, China

zoule@hfuu.edu.cn

² Institute of Applied Optimization, School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, Anhui, China

³ School of Energy Materials and Chemical Engineering, Hefei University, Hefei 230601, Anhui, China

Abstract. Handwritten chemical equations recognition is one of the important research directions of optical character recognition (OCR) and text recognition technology, which is widely used in life. Although the mainstream deep learning text recognition model can get good recognition results, the number of parameters of the model is too large to be carried on some portable devices. We develop a new lightweight network model (LCRNN) based on the CRNN model for handwritten chemical equations recognition. Firstly, in the convolutional layer of the LCRNN model, we propose a new MobileNetV3 (MobileNetV3M) to reduce number of the model parameters. The MobileNetV3M changed the original down-sampling method to max-pooling, so it can extract more critical information. Secondly, we use the BiGRU model in the recurrent layer. Finally, a new chemical equations encoding method is proposed, which can change the two-dimensional chemical equation into one-dimensional chemical equation encoding, so as to facilitate handwritten chemical equations recognition. The experiments demonstrate that the character precision of the LCRNN model is 2.1% lower than the CRNN model, but the number of parameters is significantly reduced.

Keywords: Text recognition · Lightweight network · Chemical equations · CRNN · MobileNet

1 Introduction

Text recognition is a vital research direction of optical character recognition (OCR). Handwritten chemical equations recognition and is one kind of the most important text recognition methods. Handwritten chemical equations recognition has a broad prospect in the field of education. For example, traditional text recognition [1] is to segment a text image into a single character, and then recognize the single character and combine them. However, due to the unfixed text length and the difficulty of segmentation, the

recognition accuracy is not high. Therefore, the method proposed by Liu et al. [2] used the combination of CNN and RNN to identify Chemical formulae. After combining CNN with RNN, the text recognition model can recognize the text data, whose length is unfixed more quickly and accurately.

The current mainstream text recognition model based on deep learning often combines CNN with RNN. One of them is a Convolutional Recurrent Neural Network (CRNN) [3]. The CRNN model has three layers. They are the convolutional layer, recurrent layer, and transcription layer, respectively. The convolutional layer of CRNN uses the Visual Geometry Group (VGG) [4] to extract image features and obtain the feature sequences. The recurrent layer of CRNN uses Bi-directional Long Short Term Memory (BiLSTM) [5] to learn the context information of feature sequences. The transcription layer of CRNN uses Connectionist Temporal Classification (CTC) [6] to solve the problem of text alignment. But VGG and BiLSTM networks have a large number of parameters, and they are not very friendly to some lightweight devices. Therefore, scholars carry out lightweight networks in two aspects: convolutional neural networks (such as VGG) and recurrent neural networks (such as BiLSTM).

Iandola et al. first proposed the lightweight network SqueezeNet [7]. The network SqueezeNet replaces the 3×3 convolution kernel of ordinary convolution with a 1×1 convolution kernel. In addition, the lower sampling was placed in the later stage of the network, which reduced the amount of computation but increased the amount of network computation. Amir et al. proposed a SqueezeNext [8], which is an upgraded version of SqueezeNet, and converted the $K \times K$ convolution into $1 \times K$ convolution and $K \times 1$ convolution to reduce the number of parameters. Howard et al. [9] proposed depthwise separable convolution in MobileNetV1 model. The depthwise separable convolution divided convolution into two steps: depthwise convolution and pointwise convolution. After the channels are convolved one by one, then the convolution kernel of 1×1 is used for point-by-point convolution. In order to improve the recognition accuracy of the convolutional neural network, the MobileNetV2 [10] network was proposed by Sandler et al.. The MobileNetV2 network lifts the convolution dimension, and performs depthwise convolution in high-dimensional space through 1×1 Convolution and then reduces the dimension. Howard et al. MobileNetV3 [11] proposed by adding the attention mechanism SEBlock [12] on the basis of MobileNetV2 to pay attention to the important information of each channel. In addition to the method of depthwise separable convolution, Han et al. proposed GhostNet [13] by using GhostModule to build an efficient network structure.

In the aspect of the recurrent neural network, recurrent neural network (RNN) is widely used in natural language processing. The mainstream Recurrent Neural Network (RNN) includes Long short-term Memory (LSTM) [5], Gate Recurrent Unit (GRU) [14], etc. LSTM model adds memory gate and forgetting gate on the basis of ordinary recurrent neural network (RNN), which helps neural network remember important information and forget useless information. GRU is an improved neural network based on LSTM. GRU has the same recognition accuracy as LSTM, but the calculation is simpler than LSTM, with fewer parameters, and saves more time and computation power.

CRNN [3] has a good advantage in text recognition, but its convolutional layers and recurrent layers consume large computing power and time. In order to better identify handwritten chemical equations, we propose a LCRNN model. LCRNN also has three layers: the convolutional layer, the recurrent layer and the transcription layer respectively. In order to extract better features, we propose a MobileNetV3M in the convolutional layer based on the MobileNetV3, which changes the original downsampling method to max-pooling and it can extract image features more accurately. We use the BiGRU to learn context information in the recurrent layer, the number of BiGRU parameters is significantly reduced. In addition, we construct a new encoding method for the chemical equations in this paper, which transforms two-dimensional chemical equations into one-dimensional text sequences for recognition. The chemical equations can be better recognized, and then they were reverted to the original two-dimensional chemical equations after recognition.

2 The Proposed LCRNN Model

We demonstrate the proposed LCRNN model in this section. The proposed model is a lightweight network based on the CRNN network structure. According to the characteristics of handwritten chemical equation recognition, different layers in the network are improved to solve the problem of the handwritten chemical equations. In Subsect. 2.1, the overall structure of the text recognition model LCRNN is introduced. In Subsect. 2.2, the convolutional layer of LCRNN, and lightweight convolutional neural networks MobileNetV3M are introduced. In Subsect. 2.3, the recurrent layer is introduced and two mainstream recurrent neural networks are proposed. The transcription layer and loss function are presented in Subsect. 2.4.

2.1 The Structure of the LCRNN Model

The current mainstream text recognition network CRNN consists of three parts, namely, the convolutional layer, the recurrent layer and the transcription layer. In the convolutional layer, VGG is used to extract image features to obtain feature sequences. In the recurrent layer, BiLSTM is used to learn context information; In the transcription layer, CTC loss function is used to output final results. In the convolutional layer, the CRNN model used VGG as the feature extraction network, which consists of 13 convolutions and 5 max-pooling. A large number of convolution is used, and a large number of parameters are required. It is not convenient to be loaded on lightweight devices. Therefore, we present a lightweight MobileNetV3M in the convolutional layer in Subsect. 2.2. In the recurrent layer, CRNN uses bi-directional Long short-term Memory (BiLSTM) to learn contextual information. In order to be more lightweight, the proposed LCRNN model uses the lightweight recurrent neural network BiGRU in the recurrent layer, which is introduced in Subsect. 2.3. The LCRNN model uses CRNN's transcription layer to convert the feature sequence into the output text and calculate the loss function. The transcription layer and loss function are introduced in Subsect. 2.4.

2.2 Convolutional Layer

The convolutional neural network used in the convolutional layer of LCRNN includes five downsamples and obtains feature sequences with dimension C. However, in order to better adapt to the characteristics of the input image, that is, the width of the input image is much larger than the height, the convolutional neural network uses the 2×2 method in the first two downsampling, and the 2×1 approach in the last three downsampling. As shown in Fig. 1, the height will be sampled down to 1/32 of the original. The width will be sampled down to 1/4 of the original, which can retain as much text information as possible, and make the recognition more accurate. Similarly, the downsampling of VGG, MobileNetV3, and MobileNetV3M also use the above method.

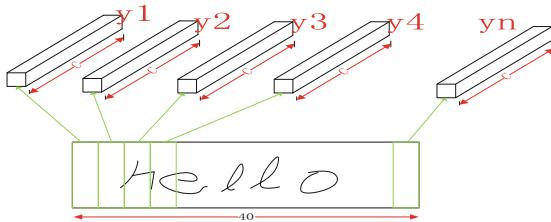


Fig. 1. Special downsampling

Next, the convolutional neural networks VGG, MobileNetV3 and MobileNetV3M in the convolutional layer are introduced respectively.

2.2.1 VGG

VGG is a very famous convolutional neural network, which has a large number of applications in the field of image classification. VGG consists of 13 convolutional layers and 5 max-pooling. However, VGG consumes more computing resources and uses multiple full-connection layers, resulting in a large number of parameters and larger storage space.

2.2.2 MobileNetV3 and MobileNetV3M

MobileNetV3 consists the Block structure and SEblock structure (SEBottleneck). This structure firstly improves the dimension through 1×1 convolution, then carries out depthwise separable convolution, and finally reduces the dimension through 1×1 convolution. In order to further improve the accuracy of model recognition, the attention mechanism of SEBlock is added to the depth separable convolution. In addition, residual (Res) structure is adopted to add the input and output directly, thus deepening the network depth. In the traditional MobileNetV3 downsampling, the convolution stride is set to 2 for downsampling, which means that downsampling is carried out in an average way, and important information characteristics in the text cannot be well remembered, just like remembering a person is to remember him by remembering his important features, rather than remembering his average characteristics. Therefore, adopting the average

method will remove some important characteristics in the text, and cannot extract the information in the text better.

Therefore, in order to better extract image features, the step size of convolution is set to 1. Then the max-pooling of 2×1 is adopted, because the max-pooling can retain the key information of the text to the maximum extent. Here, the 2×1 special down-sampling pooling is used to make more information input to the recurrent neural network (RNN), so as to improve the accuracy of recognition further. The network structure diagrams of MobileNetV3 and MobileNetV3M are shown in Fig. 2, respectively. On the left, the first downsample of MobileNetV3 is max-pooling, the second is performed by setting the stride to 2×2 , and the next three downsamples are set to 2×1 . On the right, MobileNetV3M uses max-pooling for the first downsampling, 2×1 max-pooling for the second downsampling, and sets the stride to 1×1 for the next three downsampling, and then uses 2×1 max-pooling for the next three downsampling.

MobileNetV3						
	name	inC	midC	outC	kernel	stride
	nn.Conv2d	1	16	16	3	1
						padding:1
		nn.BatchNorm2d				
		HardSwish+MaxPooling(2)				
1	SEBottleneck	16	16	16	3	2
		use SE				
		ReLU				
2	SEBottleneck	16	72	24	3	(2, 1)
		ReLU				
3	SEBottleneck	24	88	24	3	1
		ReLU				
4	SEBottleneck	24	96	40	5	(2, 1)
		use SE				
		HardSwish				
5	SEBottleneck	40	240	40	5	1
		use SE				
		HardSwish				
6	SEBottleneck	40	240	40	5	1
		use SE				
		HardSwish				
7	SEBottleneck	40	120	48	5	1
		use SE				
		HardSwish				
8	SEBottleneck	40	144	48	5	1
		use SE				
		HardSwish				
9	SEBottleneck	40	288	96	5	(2, 1)
		use SE				
		HardSwish				
10	SEBottleneck	96	576	96	5	1
		use SE				
		HardSwish				
11	SEBottleneck	96	576	96	5	1
		use SE				
		HardSwish				

MobileNetV3M						
	name	inC	midC	outC	kernel	stride
	nn.Conv2d	1	16	16	3	1
						padding:1
		nn.BatchNorm2d				
		HardSwish+MaxPooling(2)				
1	SEBottleneck	16	16	16	3	1
		use SE				
		ReLU+MaxPooling(2, 2)				
2	SEBottleneck	16	72	24	3	1
		ReLU+MaxPooling(2, 1)				
3	SEBottleneck	24	88	24	3	1
		ReLU				
4	SEBottleneck	24	96	40	5	1
		use SE				
		HardSwish+MaxPooling(2, 1)				
5	SEBottleneck	40	240	40	5	1
		use SE				
		HardSwish				
6	SEBottleneck	40	240	40	5	1
		use SE				
		HardSwish				
7	SEBottleneck	40	120	48	5	1
		use SE				
		HardSwish				
8	SEBottleneck	40	144	48	5	1
		use SE				
		HardSwish				
9	SEBottleneck	40	288	96	5	(2, 1)
		use SE				
		HardSwish+MaxPooling(2, 1)				
10	SEBottleneck	96	576	96	5	1
		use SE				
		HardSwish				
11	SEBottleneck	96	576	96	5	1
		use SE				
		HardSwish				

(a) MobileNetV3

(b) MobileNetV3M

Fig. 2. The network structure of MobileNetV3 and MobileNetV3M

2.3 Recurrent Layer

The recurrent layer needs to learn the context information of the image feature sequence, but in the process of text recognition, it needs to know not only the information behind the text, but also the information in front of the text. So bi-directional Recurrent Neural Networks are used in Network design. The classical Network are bi-directional LSTM (BiLSTM) and bi-directional GRU (BiGRU). The differences between the two will be detailed below.

BiLSTM, called Bi-directional Long-Short Term Memory, is able to remember important things and forget unimportant things by adding two gates (Memory gates and Forget gates).

Bi-directional Gated Recurrent Unit (BiGRU) is an effective variant of the BiLSTM network. It has a simpler structure and better effect than the BiLSTM network. The BiGRU uses only two gates, which combine the forget gate and the input gate into a single update gate. There was also a mix of cell and hidden states, among other changes. The BiGRU model is simpler than the standard BiLSTM model, with fewer parameters and faster training speed.

2.4 Transcription Layer and Loss Function

The transcription layer adopts the Connectionist Temporal Classification (CTC) [6], which converts the prediction results of each frame output by the recurrent layer into a tag sequence for output, so that the output sequence is aligned with the tag sequence. The probability formula is defined as follows:

$$P(L|y) = \sum_{\pi: B(\pi)=L} p(\pi|y) \quad (1)$$

where $y = y_1, y_2, \dots, y_T$ is the input sequence, T is the length of the sequence, L contains all labeled characters and a blank label in the task. Mapping function $B(\pi) = L$ will represent the mapping of the sequence π to the sequence L . $p(\pi|y) = \prod_{t=1}^T y_t^{\pi_t}$, and $y_t^{\pi_t}$ is the probability of having label π_t at moment t .

The loss function is defined as follows:

$$O = - \sum_{I_i, L_i} \log P(L_i|y_i) \quad (2)$$

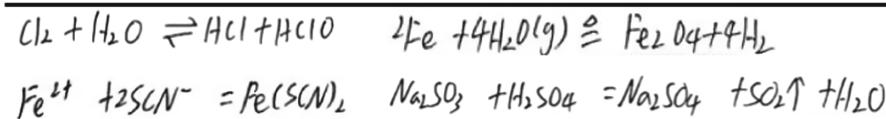
where $X = \{I_i, L_i\}$ denotes the training set, I_i denotes the training image, L_i denotes the true label sequence, and y_i denotes the output of the recurrent layer.

In summary, based on CRNN, MobileNetV3M, and BiGRU, CTC loss function, we propose the lightweight model LCRNN. The LCRNN network in this paper enables end-to-end character-free segmentation training and prediction, thus avoiding the extra work involved in segmenting each character position to be recognized.

3 Data Sets

3.1 Collection Set of Data Sets

The handwritten chemical equation data set used in this paper is constructed as follows: Firstly, 170 common used chemical equations and ion equations were collected. Secondly, each chemical equation was manually written 10 times to obtain 1700 images. Thirdly, 1700 handwritten chemical equations are then blurred by Gaussian with radius 1 to obtain blur 1700 images. There are 3400 original images and blur images. Finally, 3400 images were divided into the training set and the test set in the form of 9:1, among which 3060 images were used as training set and 340 images were used as test set. For the convenience of subsequent training and recognition, we pre-process the image. The pre-process of data sets includes the two steps: (1) Scale the image in equal proportions and change the height to 32; (2) Set the picture as the intensity image. Some examples of chemical equation images are shown in Table 1.

Table 1. The handwritten chemical equation

3.2 Labels of Data Sets

The handwriting recognition model often focus on English characters, numbers, and characters, and have a good recognition performance in these areas. However, these identifications are limited to a one-dimensional space and chemical equations are two-dimensional structures, i.e., there are superscripts and subscripts, which make identification difficult. So the recognition effectiveness of handwritten chemical formulas is still a difficult task. Handwritten chemical equation datasets is a special text data. They are different from regular text recognition datasets tags. The chemical equation includes superscript and subscript, it is a two-dimensional label, so special label methods are needed to make them easy to identify.

We classify the labels in the handwritten chemical equation datasets into four categories: the English alphabets, digital, special, and ionic symbols. We describe the specific labels of these four categories as follows.

- (1) English alphabets: a–z, A–Z
- (2) Digital: 0–9
- (3) Special symbols

Subscripts exist in chemical equations, so we construct a new label method, which can change the two-dimensional chemical equations into one-dimensional chemical equation encoding. We use an underscore “_” to indicate the subscript, and Table 2 shows examples of the use of underscores. In addition to subscripts, there are special reaction symbols in the chemical equations. We define a new method for the special reaction symbols, the different special reaction symbols are shown in Table 3.

- (4) Ionic symbols

In the handwritten chemical equation, there are also ionic equations with superscripts, we define a new method for the ionic symbols. The ionic symbols of them are shown in Tables 4, and 5. Finally, we give the example of encoding different chemical equations in Table 6.

Table 2. Encoding of chemical expressions

Expression	Code	Expression	Code	Expression	Code
CaCl_2	CaCl_2	H_2SO_4	H_2SO_4	$\text{Ca}(\text{ClO})_2$	$\text{Ca}(\text{ClO})_2$

Table 3. Encoding of special symbols in chemical equations

symbol	code	symbol	code	symbol	code	symbol	code	symbol	code
↑	^	↓	!	↔	↔	→	→	Δ	~
<u>点燃</u>	*=	<u>通电</u>	\&=	<u>高温</u>	\\$=	<u>点燃</u> →	*>	<u>光</u>	\@=

Table 4. Ionic Equations

+	2+	3+	4+	5+	6+	7+	-	2-	3-	4-	5-	6-	7-
---	----	----	----	----	----	----	---	----	----	----	----	----	----

Table 5. Encoding of Ionic expressions

Ion expression	Code	Ion expression	Code	Ion expression	Code
2OH⁻	2OH −	HCO₂⁻	HCO_2 −	Mg²⁺	Mg 2+

Table 6. Examples of chemical equations after coding

Chemical equation	The encoded expression
$2\text{Na} + 2\text{H}_2\text{O} + \text{CuSO}_4 = \text{Cu}(\text{OH})_2\downarrow + \text{Na}_2\text{SO}_4 + \text{H}_2\uparrow$	$2\text{Na} + 2\text{H}_2\text{O} + \text{CuSO}_4 = \text{Cu}(\text{OH})_2! + \text{Na}_2\text{SO}_4 + \text{H}_2^{\wedge}$
$\text{Ca}^{2+} + 2\text{OH}^- + 2\text{HCO}_2^- = \text{CaCO}_2\downarrow + 2\text{H}_2\text{O} + \text{CO}_2^{2-}$	$\text{Ca}2+ + 2\text{OH} - + 2\text{HCO}_2 - = \text{CaCO}_2! + 2\text{H}_2\text{O} + \text{CO}_2 2-$

There are four steps when converting a one-dimensional coding to a two-dimensional Chemical equation. (1) Defining reaction symbols be “=” or two consecutive characters following the character “\”. We divide a chemical equation into two parts by using a reaction symbol. (2) Processing ion symbols, that is, if the first character after “|” is “+”, then converts it to a “*”, to prevent the “+” confusion. (3) Using the “+” to divide into several chemical expressions. (4) Converting each chemical expression.

For example, for one-dimensional coding “ $\text{Ca}2+ + 2\text{OH}|- + 2\text{HCO}_2| - = \text{CaCO}_2! + 2\text{H}_2\text{O} + \text{CO}_2|2-$ ”. (1) “=” is the reaction symbol, it is divided into “ $\text{Ca}2+ + 2\text{OH}|- + 2\text{HCO}_2| -$ ” and “ $\text{CaCO}_2! + 2\text{H}_2\text{O} + \text{CO}_2|2-$ ”. (2) Convert “+” after the ion symbol “|” to “*”, getting “ $\text{Ca}2* + 2\text{OH}|- + 2\text{HCO}_2| -$ ” and “ $\text{CaCO}_2! + 2\text{H}_2\text{O} + \text{CO}_2|2-$ ”. (3) Divided “+” to get $\text{Ca}2*$, $2\text{OH}|-$, $2\text{HCO}_2| -$, $\text{CaCO}_2!$, $2\text{H}_2\text{O}$, $\text{CO}_2|2-$. (4) After converting, we can get Ca^{2+} , 2OH^- , 2HCO_2^- , $\text{CaCO}_2\downarrow$, $2\text{H}_2\text{O}$, CO_2^{2-} .

We give another example to express the process of the one-dimensional coding. For one-dimensional coding “ $2\text{CH}_2\text{CH}_2\text{OH} + 2\text{Na}\rightarrow 2\text{CH}_2\text{CH}_2\text{ONa} + \text{H}_2\wedge$ ”. (1)

Taking the reaction symbol “ \rightarrow ” and see them as a whole, and dividing them into “ $2\text{CH}_2\text{CH}_2\text{OH} + 2\text{Na}$ ” and “ $2\text{CH}_2\text{CH}_2\text{ONa} + \text{H}_2\uparrow$ ”. (2) Because there’s no ion sign “ l^- ”, we don’t have to deal with it. (3) Using “ t ” to divide them into $2\text{CH}_2\text{CH}_2\text{OH}$, 2Na , $2\text{CH}_2\text{CH}_2\text{ONa}$, $\text{H}_2\uparrow$. (4) After converting, we can get $2\text{CH}_2\text{CH}_2\text{OH}$, 2Na , $2\text{CH}_2\text{CH}_2\text{ONa}$, $\text{H}_2\uparrow$.

4 Experiment Results and Analysis

In this section, we will give some experiments to demonstrate the effectiveness of the proposed LCRNN model. For all the experiments, CPU is Intel (R) Core (TM) I5-9500 CPU @ 3.00 GHz, the GPU is Nvidia GTX 1080 Ti 11 GB video memory capacity.

4.1 Evaluation Indicators

(1) Average Precision (AP)

$$\text{acc} = \frac{\text{correct_num}}{\text{all}} \quad (3)$$

where `correct_num` is the number of exactly correct string and `all` is the total number.

(2) Average Precision of Character (APc)

The comparison of strings is different from the classification, and the difference of one character cannot erase the credit of other characters being the same, so this paper mainly uses APc as a reference basis.

It is calculated using the `SequenceMatcher` method in python’s `difflib` library with the following code: “`difflib.SequenceMatcher(None, str1, str2).ratio()`”.

The formula is calculated by comparing the similarity of strings a, string b and finding the same substrings “ s_1s_2,\dots,s_n ” of the same a and b. And “ s_1s_2,\dots,s_n ” satisfy the following 2 constraints.

① $\text{end}(s_i) < \text{start}(s_{i+1})$ $1 \leq i \leq n - 1$, where “`start`” is the start index of the string and “`end`” is the end index of the string.

② $\sum_{i=1}^n \mathcal{L}(s_i) \leq \mathcal{L}(a)$ and $\sum_{i=1}^n \mathcal{L}(s_i) \leq \mathcal{L}(b)$.

The formula for calculating the similarity ratio between string a and string b is as follows.

$$\text{ratio} = 2 \times \frac{\sum_{i=1}^n \mathcal{L}(s_i)}{\mathcal{L}(a) + \mathcal{L}(b)}, \text{L indicates the length of the string.} \quad (4)$$

We give an example: let the string $a = \text{"helloworldxxx"}$ and the string $b = \text{"myworld-mxxx"}$. Then the length of string a is 13 and the length of string b is 11. The identical substrings of string a and string b are $s_1 = \text{"world"}$ and $s_2 = \text{"xxx"}$, where the length of s_1 is 5 and the length of s_2 is 3. Then the similarity ratio calculation formula of string a and string b is as follows: $\text{ratio} = 2 \times \frac{5+3}{13+11} \approx 0.667$.

4.2 Model Performance Comparison

In order to verify the effect of max-pooling on down-sampling, we will use max-pooling in different down-sampling. Dn represents the n-th downsampling stage, S represents down-sampling through stride, and M represents max-pooling. In addition, (2,2) represents the 2×2 down-sampling method, and (2,1) represents the 2×1 down-sampling method. The first line represents the original MobileNetV3 which shown in Fig. 2 (a). From the second line, different down-sampling is replaced by max-pooling. With the increase of max-pooling, the accuracy increases. When all the down-sampling methods are changed into max-pooling, the APc is 97.512%. So the text uses the last line as the MobileNetV3M network structure. We think this is because the handwritten chemical equation data set used in this paper is black with a white background. The background is simple. Therefore, max-pooling can better highlight the characteristics of images (Table 7).

Table 7. Accuracy at different stages using different downsampling methods

D1	D2	D3	D4	D5	AP (%)	APc (%)	Note
M(2,2)	S (2,2)	S (2,1)	S (2,1)	S (2,1)	43.5	90.002	MobileNetV3
M(2,2)	M(2,2)	S (2,1)	S (2,1)	S (2,1)	49.4	93.611	
M(2,2)	M(2,2)	S (2,1)	S (2,1)	M(2,1)	50.3	93.468	
M(2,2)	M(2,2)	S (2,1)	M(2,1)	M(2,1)	57.6	95.383	
M(2,2)	M(2,2)	M(2,1)	M(2,1)	M(2,1)	74.1	97.512	MobileNetV3M

In order to demonstrate the effectiveness of the LCRNN model, different lightweight models were used to conduct experiments on chemical equation data sets, and the results were compared.

Table 8. Accuracy of text recognition models using BiGRU at the Recurrent Layer on a chemical equation dataset

Convolutional layer	Recurrnnet layer	Weights (KB)	AP (%)	APc (%)
VGG	BiGRU	32114	95.9	99.659
MobileNetV3	BiGRU	17285	43.5	90.002
MobileNetV3M	BiGRU	17320	74.1	97.512

In the first row, VGG is used in the convolutional layer, and BiGRU is used in the recurrent layer. The second row uses MobileNetV3 in the convolutional layer and BiGRU in the recurrent layer, and the third row uses MobileNetV3M in the convolutional layer and BiGRU in the recurrent layer. As shown in Table 8, when BiGRU is used in the recurrent layer, the APc of VGG, MobileNetV3 and MobileNetV3M in the convolutional

layer is 99.659%, 90.002% and 97.512% respectively. It can be seen that MobileNetV3M has a significant improvement over MobileNetV3, but the number of parameters is significantly reduced.

Table 9. Accuracy of text recognition models using BiLSTM at the Recurrent Layer on a chemical equation dataset

Convolutional layer	Recurrnet layer	Weights (KB)	AP (%)	APc (%)
VGG	BiLSTM	34682	96.5	99.614
MobileNetV3	BiLSTM	19851	54.7	92.889
MobileNetV3M	BiLSTM	19888	72.9	97.691

In the first row, VGG is used in the convolutional layer and BiLSTM is used in the recurrent layer. The second row uses MobileNetV3 in the convolutional layer and BiLSTM in the recurrent layer, and the third row uses MobileNetV3M in the convolutional layer, and BiLSTM in the recurrent layer. As shown in Table 9, when BiLSTM is used in the recurrent layer, the APc of VGG, MobileNetV3, and MobileNetV3M is 99.614%, 92.889% and 97.691%, respectively. Similar to the recurrent layer using BiGRU. MobileNetV3M has a significant improvement in accuracy compared to MobileNetV3, because MobileNetV3M changes the down-sampling mode to max-pooling, which means MobileNetV3M can extract the key features of characters and improve the accuracy of recognition by these key features. And the number of parameters significantly reduced.

Table 10. Comparison of accuracy between CRNN and LCRNN

Network	Weights (KB)	AP (%)	APc (%)
CRNN(VGG + BiLSTM)	34682	96.5	99.614
LCRNN(MobileNetV3M + BiGRU)	17320	74.1	97.512

The APc of LCRNN is 2.102% different from that of CRNN, but the number of parameters is half that of CRNN. Therefore, LCRNN can reduce the number of model parameters while ensuring the accuracy of model recognition. So MobileNetV3M in the convolutional layer and BiGRU in the recurrent layer is the better choice in the lightweight handwritten chemical equations recognition model (Table 10).

5 Conclusion

In this paper, we demonstrate a lightweight CRNN model (LCRNN). The LCRNN model is a lightweight improvement on the convolutional layer and recurrent layer in the CRNN model, it greatly reducing the parameters of the original CRNN model. The convolutional layer of LCRNN uses MobileNetV3M. MobileNetV3M can preserve the characteristics of the image to the maximum extent. The LCRNN's recurrent layer uses BiGRU to further reduce the number of parameters in the model. We also construct a new format for encoding chemical equations, it transforms two-dimensional chemical equations into one-dimensional text sequences for recognition. Experiments are given to show that LCRNN was close to the APc of CRNN, but the number of parameters was significantly reduced. In the future, some other lightweight models will be used to further reduce the number of parameters in the model and ensure the accuracy of recognition.

Acknowledgment. This work was supported by the grant of the Hefei College Postgraduate Innovation and Entrepreneurship Project, No. 21YCXL16, the grant of Anhui Provincial Natural Science Foundation, Nos. 1908085MF184, 1908085QF285, the grant of Scientific Research and Talent Development Foundation of the Hefei University, No. 18-19RC26, in part by the grant of Key Generic Technology Research and Development Project of Hefei, No. 2021GJ030, the Key Research Plan of Anhui Province, Nos. 202104d07020006, 2022k07020011.

References

1. Yang, J., Shi, G., Wang, K., et al.: A study of on-line handwritten chemical expressions recognition. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
2. Liu, X., Zhang, T., Yu, X.: An end-to-end trainable system for offline handwritten chemical formulae recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 577–582 (2019)
3. Shi, B., Xiang, B., Cong, Y.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(11), 2298–2304 (2016)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *Proc. Int. Conf. Learn. Represent.* (2015)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
6. Graves, A.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proc. Int. Conf. Mach. Learn.* (2006)
7. Iandola, F., Han, S., Moskewicz, M., et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *Comput. Vision Pattern Recognit.* (2016)
8. Gholami, A., Kwon, K., Wu, B., et al.: SqueezeNext: hardware-aware neural network design. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018)
9. Howard, A.G., Zhu, M., Chen, B., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017)
10. Sandler, M., Howard, A., Zhu, M., et al.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

11. Howard, A., Sandler, M., Chen, B., et al.: Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2020)
12. Hu, J., Shen, L., Sun, G., et al.: Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 2011–2023, 2020 (2017)
13. Han, K., Wang, Y., Tian, Q., et al.: GhostNet: more features from cheap operations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
14. Chung, J., Gulcehre, C., Cho, K., et al.: Gated feedback recurrent neural networks. In: International Conference on Machine Learning, pp. 2067–2075 (2015)



Illumination Invariant Face Recognition Using Directional Gradient Maps

Guang Yi Chen¹(✉), Wenfang Xie², and Adam Krzyzak¹

¹ Department of Computer Science and Software Engineering, Concordia University, Montreal,
QC H3G 1M8, Canada

guangyi_chen@hotmail.com, krzyzak@cse.concordia.ca

² Department of Mechanical and Industrial Engineering, Concordia University, Montreal,
QC H3G 1M8, Canada

wfxie@me.concordia.ca

Abstract. Face recognition under varying illumination is a difficult problem in many real-life applications. Various methods have been developed in the literature to deal with this problem. In this paper, we propose a new method for face recognition under varying lighting conditions. Our method calculates the difference of pixel intensity values along a direction and then divides the difference by the mean intensity value of a small region centered at the pixel to generate a map for this direction. It can be shown that this directional gradient map is approximately illumination invariant. Our experiments show that our proposed method compares favourably to existing methods especially under the noisy environments. For the noise-free Extended Yale Face Database B, we obtained 95.4% correct recognition rate, and for the noise-free CMU-PIE face database, we achieved 100% correct recognition rate.

Keywords: Face recognition · Invariant features · Illumination invariant · Collaborative representation-based classifier (CRC)

1 Introduction

Face Recognition is an especially important research topic for recognizing human faces from images or videos. Even for the same person, face images appear differently because of lighting conditions, expression, pose, occlusion, and so forth. Face recognition has such wide applications in information security, smart cards, entertainment, law enforcement and surveillance. Nevertheless, existing methods may fail to perform well in these conditions due to illumination difference.

There are several existing face recognition methods in the literature, developed for varying illumination environments. Lee et al. [1] studied illumination invariant face recognition by arranging physical lighting. Du and Ward [2] were the first to develop an adaptive region-based image enhancement method for face recognition by means of 2D wavelet transform. Nevertheless, this method introduces defects along the boundary of different image regions. Chen et al. [3] first performed logarithm transform to the face

images and then conduct discrete cosine transform (DCT) to the logarithm images (LOG-DCT). They set the low-resolution DCT coefficients to zero and then perform inverse DCT to obtain illumination invariant face maps for classification. Chen et al. [4] proposed a logarithmic total variation model (LTV) for face recognition under variable lighting condition. However, it is relatively complex than other methods because it needs to solve differential equations. Ahonen et al. [5] developed local binary patterns (LBP) for face recognition. They divided the face images into several regions, extracted the LBP feature distributions, and then concatenated into a feature vector as a face descriptor. Zhang et al. [6] investigated illumination invariant face recognition using gradient faces. However, these gradient faces are relatively sensitive to noise. Lai et al. [7] developed a new method for illumination invariant face recognition by defining multiscale logarithm difference edge-maps. They utilized a logarithm difference model, which eliminates light intensity from pixels within a small neighbourhood. It was claimed that this method is better than LOG-DCT and LTV. Shah et al. [8] proposed a robust face recognition technique under varying illumination condition. They transformed pixels from non-illuminated side to illuminated side and reported that their proposed method produced better results than other existing methods compared in their paper.

In this paper, we propose a novel method for face recognition by extracting directional gradient maps from the face images. We use collaborative representation-based classifier (CRC) to classify face images. It has been tested on two publicly available databases and has been found that the proposed method can handle the problem of variations in facial images due to variations in illumination in a robust manner. In addition, our new method is better than gradient faces under noisy environment for most of the test cases conducted in this paper.

The organization of this paper is as follows. Section 2 introduces a new method to extract illumination invariant face maps by taking the directional gradient. Section 3 conducts some experiments to show the effectiveness of the proposed method. Finally, Section 4 concludes the paper and proposes future research directions.

2 Proposed Study

According to the Lambertian reflectance theory [9], the intensity image can be modeled as.

$$I(x, y) = R(x, y)L(x, y) \quad (1)$$

where R is the reflectance and L is the illumination. Because R depends only on the surface material of the subject, it is the intrinsic representation of a face image. Like gradient faces, we propose to extract directional gradient maps along any directions. In this paper, we only consider two directions with angles 45° and 135° , respectively. For any pixel $I(x, y)$ in a face image, we define two direction maps:

$$Map_{45} = \arctan\left(\frac{Dif_{45}}{Avg_{45}}\right), \quad (2)$$

$$Map_{135} = \arctan\left(\frac{Dif_{135}}{Avg_{135}}\right), \quad (3)$$

where

$$\begin{aligned} Dif_{45}(x, y) &= I(x+1, y+1) + I(x+2, y+2) \\ &\quad - I(x-1, y-1) - I(x-2, y-2) \end{aligned} \quad (4)$$

$$Avg_{45}(x, y) = \frac{1}{5 \times 5} \sum_{i=x-2}^{x+2} \sum_{j=y-2}^{y+2} I(i, j) \quad (5)$$

$$\begin{aligned} Dif_{135} &= I(x-1, y+1) + I(x-2, y+2) \\ &\quad - I(x+1, y-1) - I(x+2, y-2) \end{aligned} \quad (6)$$

$$Avg_{135} = \frac{1}{5 \times 5} \sum_{i=x-2}^{x+2} \sum_{j=y-2}^{y+2} I(i, j) \quad (7)$$

Theorem 1. The direction maps $Map_{45}(x, y)$ and $Map_{135}(x, y)$ defined above are approximately illumination invariant.

Proof: Since, within a 5×5 neighbourhood of pixel $I(x, y)$, the illumination values are approximately the same, we can write:

$$\begin{aligned} Dif_{45}(x, y) &\cong (R(x+1, y+1) + R(x+2, y+2) \\ &\quad - R(x-1, y-1) - R(x-2, y-2))L(x, y) \end{aligned} \quad (8)$$

$$Avg_{45}(x, y) \cong \frac{L(x, y)}{5 \times 5} \sum_{i=x-2}^{x+2} \sum_{j=y-2}^{y+2} R(i, j) \quad (9)$$

$$\begin{aligned} Dif_{135}(x, y) &\cong (R(x-1, y+1) + R(x-2, y+2) \\ &\quad - R(x+1, y-1) - R(x+2, y-2))L(x, y) \end{aligned} \quad (10)$$

$$Avg_{135}(x, y) \cong \frac{L(x, y)}{5 \times 5} \sum_{i=x-2}^{x+2} \sum_{j=y-2}^{y+2} R(i, j) \quad (11)$$

Hence, the illumination term $L(x, y)$ will disappear in the direction maps due to the division Dif and Avg :

$$\frac{Dif_{45}(x, y)}{Avg_{45}(x, y)} \text{ and } \frac{Dif_{135}(x, y)}{Avg_{135}(x, y)}$$

This completes the proof. □

To perform well in the noisy environment, we perform smoothing to the face images before extracting our directional gradient maps. This is the same as gradient faces [6]. We do not perform smoothing if we conduct denoising to the face images. We use the CRC to classify face images to one of the known classes. Our method is approximately invariant to illumination changes in face images, and it is better than gradient faces under noisy environment for most of the test cases.

The steps of our proposed method can be listed as follows:

- Step 1. Given an input unknown face image $I(x,y)$. Either perform image denoising to $I(x,y)$ by means of any good denoising methods or smooth the input face image $I(x,y)$ by convolving with a Gaussian kernel just like gradient face [6]:

$$I_1 = I * G(x, y, \sigma), \quad (12)$$

where $G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$ is a Gaussian kernel function with standard deviation σ .

- Step 2. Compute Map_{45} and Map_{135} from the image generated from Step 1 according to Eqs. (2) and (3), respectively.
 Step 3. Normalize $D = Map_{45}$ or $D = Map_{135}$ to have zero mean and unit variance.
 Step 4. Set $E = D^k$, where $k = 0.69$ is a constant.
 Step 5. Use the normalized Map_{45} or Map_{135} as features to classify the input face image $I(x,y)$ to one of the known classes by CRC [12].

Our new method is different from gradient faces. Gradient face of a face image $I(x,y)$ is defined as

$$G(x, y) = \arctan\left(\frac{\partial I(x, y)}{\partial y} / \frac{\partial I(x, y)}{\partial x}\right) \quad (13)$$

which is different from our definition (2) and (3). In addition, our direction maps are relatively less sensitive to noise than gradient faces as demonstrated in the next section. However, our method has similar computational complexity as the gradient faces.

3 Experimental Results

We conduct some experiments for face recognition using the Extended Yale Face Database B [1] and the CMU Pose, Illumination and Expression (PIE) (CMU-PIE) illumination face database [10]. The first database contains faces of 38 subjects in 64 diverse lighting conditions: from normal to extremely badly illuminated. There is only one ideal image for each subject. There are 2414 available images in total, and they are already aligned well. The cropped and normalized faces of size 192×168 were captured under various laboratory - controlled lighting conditions. We take one well-lighted face image as the single reference and take all the rest available $2414 - 38 = 2376$ images as test samples. The faces are divided into 5 subsets according to angles between the light source direction and the camera axis (Table 1). We scale the direction maps to 64×64 pixels as the features for this face database. The degree of variation gets higher from Subset 1 to Subset 5. Figure 1 shows the five subsets for one subject.

Table 1. The five subsets of the Extended Yale Face Database, their corresponding angles, and the number of faces in each subset.

Subsets	Angles	Number of Faces
Subset 1	$1^\circ \leq \text{angle} \leq 12^\circ$	7×38
Subset 2	$13^\circ \leq \text{angle} \leq 25^\circ$	12×38
Subset 3	$26^\circ \leq \text{angle} \leq 51^\circ$	12×38
Subset 4	$52^\circ \leq \text{angle} \leq 77^\circ$	14×38
Subset 5	$78^\circ \leq \text{angle}$	19×38

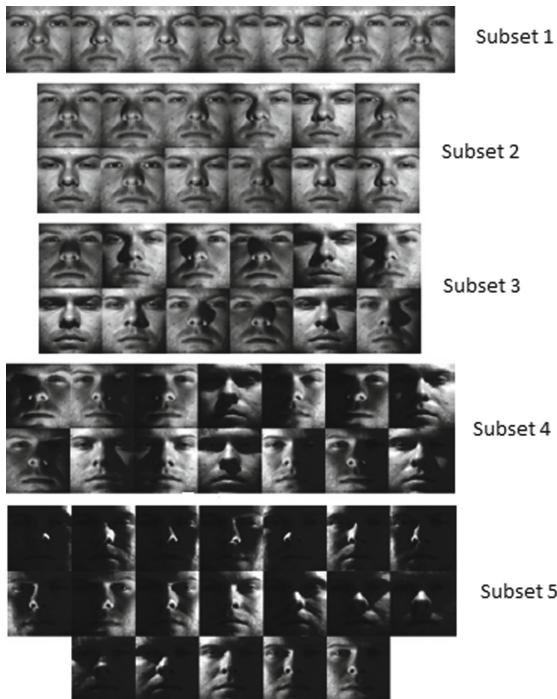


Fig. 1. The five subsets of the Extended Yale-B face database.

The second database consists of 41368 images of 68 subjects. Each subject has images captured under 13 different poses and 43 different illumination conditions and with four different expressions. We only select a subset that focuses on illumination variations on light intensity and direction in frontal view. There exist 68 subjects in each 43 images yielding a total of 2924 images. We scale the direction maps to 128×128 pixels as the features for this face database. Figure 2 shows different original face images under different lighting condition in this database.



Fig. 2. An example of the face images under different lighting condition of the CMU-PIE illumination face database.

Our correct recognition rates for $Map_{135}(x, y)$ are shown in Tables 2 and 3, where the correct recognition rate is defined as the percentage of faces that are classified to its true class. We do not report the results for $Map_{45}(x, y)$ because it is not as good as $Map_{135}(x, y)$. In Table 2, we compare our results under noise-free environment with Large and small scale [11], gradient faces [6], LTV [4], Local binary pattern [5], histogram equalization [13], and no features extraction (None). Note that we copy the values of all these compared methods except gradient faces from [11]. Our $Map_{135}(x, y)$ generates the highest recognition rates for the Extended Yale Face Database B, and it is

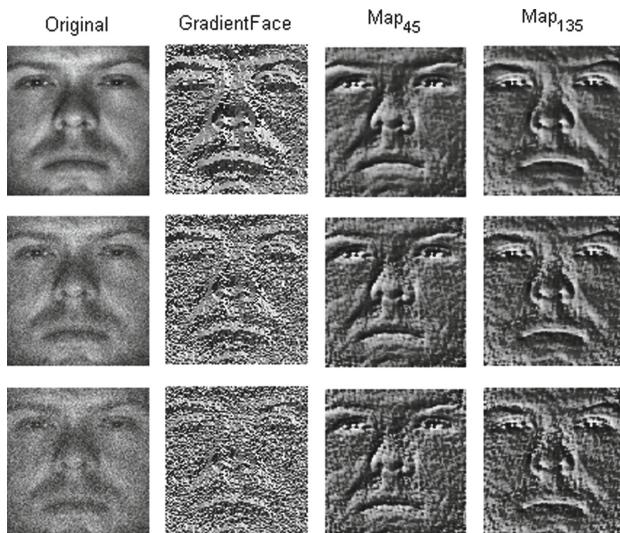


Fig. 3. A comparison between the gradient faces and our directional gradient maps: Map_{45} and Map_{135} . The first/second/third row has noise standard deviation $\sigma_n = 10/20/30$, respectively. Gradient faces are relatively less robust to Gaussian white noise than our proposed methods in this paper.

comparable to gradient faces (both 100%) for the CMU-PIE face database. Our experimental results under noisy environment (Gaussian white noise) are tabulated in Table 3. As can be seen, for the Extended Yale Face Database B, our method is better than gradient faces for all test cases no matter we use denoising as a preprocessing step or not. For the CMU-PIE database, our method is better than the gradient faces when we use denoising as a preprocessing step. However, our method is comparable with gradient faces when we do not use denoising as a preprocessing step.

Figure 3 shows a comparison between the gradient faces and our directional gradient maps: Map_{45} and Map_{135} . The first/second/third row has noise standard deviation $\sigma_n = 10/20/30$, respectively. Gradient faces are relatively less robust to noise than our proposed methods as demonstrated in this figure. Both Map_{45} and Map_{135} can eliminate noise a lot in the extracted face images, but gradient faces contain significant amount of noise.

Table 2. The correct classification rates (%) of the proposed method in this paper, and of the methods: Large and small scale [11], gradient faces [6], LTV [4], Local binary pattern [5], histogram equalization [13], and no features extraction (None). All values for these compared methods except gradient faces are taken from [11]. The best results are highlighted in bold font,

Methods	CMU-PIE	Extended Yale-B					
		Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Average
Map_{135}	100	100	100	94.5	92.4	89.9	95.4
Large and Small Scale features [11]	99.9	100	100	86.0	85.3	84.8	91.2
Gradient Faces [6]	100	98.7	100	92.5	86.3	86.6	92.8
LTV [4]	99.8	100	99.8	78.5	75.8	82.4	87.3
Local Binary Pattern [5]	75.4	100	100	62.3	10.3	6.6	55.8
Histogram Equalization [13]	42.2	99.1	94.7	43.2	12.2	15.4	52.9
None	35.1	99.6	96.7	41.1	7.4	3.2	49.6

Table 3. The correct classification rates (%) of the proposed method in this paper and of Gradient Faces for face images corrupted by Gaussian white noise. In the table, (1) means performing denoising and (2) means without denoising. Noise standard deviation (σ_n) is in the range of 5 to 40. The best results are highlighted in bold font.

Databases	Methods	Noise Standard Deviation (σ_n)							
		5	10	15	20	25	30	35	40
Extended Yale-B	$Map_{135}(1)$	91.3	88.0	84.9	82.1	79.3	77.4	75.7	74.8
	Gradient Faces (1)	87.0	83.9	81.6	79.0	77.3	76.2	75.3	74.1
	$Map_{135}(2)$	91.0	85.5	82.1	79.2	77.3	75.7	74.9	74.2
	Gradient Faces (2)	80.9	74.6	70.6	7.1	64.4	61.3	58.1	55.4

(continued)

Table 3. (*continued*)

Databases	Methods	Noise Standard Deviation (σ_n)							
		5	10	15	20	25	30	35	40
CMU-PIE	<i>Map</i> ₁₃₅ (1)	100	100	100	100	100	100	100	99.9
	Gradient Faces (1)	99.9	99.7	99.6	99.6	99.6	99.6	99.4	99.3
	<i>Map</i> ₁₃₅ (2)	100	100	100	99.9	99.6	98.9	98.0	97.7
	Gradient Faces (2)	99.9	99.9	99.7	99.7	99.6	99.4	99.2	99.0

4 Conclusions

The appearance of a human face will change drastically if the illumination varies a lot in the recorded face images. In addition, variations in lighting conditions can make face recognition an even more challenging and difficult task.

In this paper, we have proposed a novel method for face recognition in varying illumination environment. Our method extracts two direction maps with direction angles 45° and 135°, respectively. It has been shown that these two maps are invariant to illumination changes in face images. Experimental results show that our new map with 135° compares favourably to the gradient faces for both the Extended Yale Face Database B and the CMU-PIE illumination face database. Our method performs especially well under noisy environment when compared to the gradient faces.

In our future research, we would like to test our proposed method in this paper for uncontrolled outdoor lighting conditions. This is because both the Extended Yale Face Database B and the CMU-PIE illumination face database are captured under controlled indoor lighting conditions. We will choose deep learning for illumination invariant face recognition in our future research. We will try deep convolution networks (CNN) for illumination invariant face recognition. We may also compare our previous work [14] with the method proposed in this paper for illumination invariant face recognition.

References

1. Lee, K.C., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(5), 684–698 (2005)
2. Du, S., Ward, R.K.: Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions. *IEEE Trans. Circuits Syst. Video Technol.* **20**(9), 1165–1175 (2010)
3. Chen, W., Er, M. and Wu, S.: Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Trans. Syst. Man Cybern. B Cybern.* **36**(2), 458–466 (2006)
4. Chen, T., Yin, W., Zhou, X.S., Comaniciu, D., Huang, T.S.: Total variation models for variable lighting face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(9), 1519–1524 (2006)
5. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(12), 2037–2041 (2006)
6. Zhang, T., Tang, Y.Y., Fang, B., Shang, Z., Liu, X.: Face recognition under varying illumination using gradientfaces. *IEEE Trans. Image Process.* **18**(11), 2599–2606 (2009)

7. Lai, Z.R., Dai, D.Q., Ren, C.X., Huang, K.K.: Multiscale logarithm difference edgemaps for face recognition against varying lighting conditions. *IEEE Trans. Image Process.* **24**(6), 1735–1747 (2015)
8. Shah, J.H., Sharif, M., Raza, M., Murtaza, M., Ur-Rehman, S.: Robust Face Recognition Technique under Varying Illumination. *J. Appl. Res. Technol.* **13**(1), 97–105 (2015)
9. Horn, B.K.P.: *Robot Vision*, Cambridge. MIT Press, MA (1997)
10. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(12), 1615–1618 (2003)
11. Xie, X., Zheng, W., Lai, J., Yuen, P.C., Suen, C.Y.: Normalization of face illumination based on large and small-scale features. *IEEE Trans. Image Process.* **20**, 1807–1821 (2011)
12. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: which helps face recognition? In: *IEEE International Conference on Computer Vision*, pp. 471–478 (2011)
13. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. (2008)
14. Chen, G.Y., Bui, T.D., Krzyzak, A.: Illumination invariant face recognition using dual-tree complex wavelet transform in logarithm domain. *J. Electr. Eng.* **70**(2), 113–121 (2019)



News Video Description Based on Template Generation and Entity Insertion

Qiyang Yuan, Pengjun Zhai^(✉), Dulei Zheng, and Yu Fang

Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
{2030770,1810369,2111463,fangyu}@tongji.edu.cn

Abstract. News video description aims to generate a knowledge-rich description for a news video with attached text. The difficulty of this task is how to mine events and named entities from attached text using video input. Existing approaches are all based on one-stage methods and do not filter redundant contextual sentences, resulting in inaccurate descriptions. This paper proposes a two-stage approach based on template generation and entity insertion, where the first stage focuses on the generation of events and the second stage focuses on the generation of named entities such as event participants. Specifically, we first design a sentence ranker based on pre-trained models to filter video-related sentences from the attached text, then use a multimodal encoder and transformer-based decoder to generate a description template, and finally do the entity insertion using the sorted sentences to get the final description. The results show that our method exhibits strong performance on the News Video Dataset.

Keywords: News video description · Template generation · Entity insertion

1 Introduction

With the development of internet technology, videos are rapidly produced and disseminated. General video descriptions aim to produce short descriptions for videos without attached text. Previous works [1, 2] have made great progress in the general video description. However, the descriptions produced by these models are “general”, lacking domain-specific background knowledge, especially entity information. Due to the characteristics of news videos, the generated descriptions need to include time, place, person, the cause, process, and result of the event, etc. The previous methods do not work for news video descriptions.

News video description [3] enriches the generated descriptions by introducing the attached text as background knowledge. On the one hand, entity information such as names and locations can be extracted from the attached text, and on the other hand, the attached text can guide the model to generate more accurate descriptions. A video has many contextual sentences with high or low relevance. The previous approaches [3, 4] do not filter the contextual sentences and are all based on one-stage methods, which cannot make full use of the attached text, resulting in low accuracy.

This paper proposes a two-stage method based on template generation and entity insertion. Specifically, we first design a sentence ranker based on the pre-trained language model BERT [5] to sort the contextual sentences by similarity to video. Then a certain number of sentences are selected for subsequent template generation and entity insertion. In the first stage, an encoder-decoder model based on a multi-modal encoder and a transformer-based [6] decoder is used to generate description templates with entity placeholders. In the second stage, entities of the corresponding category are selected from the sorted contextual sentences. Figure 1 shows an example of our approach. The results show that our model can produce descriptions that are closer to ground truth and outperforms previous methods on multiple metrics of the news video description.

Our main contributions include:

1. We propose a sentence ranker, which can be used as a basic component for vision and text similarity calculations
2. We propose a two-stage news video description method, including template generation and entity insertion.
3. Our method exhibits strong performance on the news video dataset and demonstrates its effectiveness through contrastive and ablation experiments.

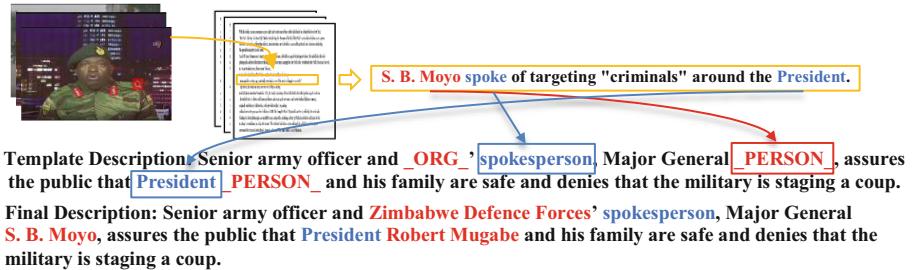


Fig. 1. An example of our approach. Given a video and its attached text, our model first uses the video to filter the text and generate a description template. Then the entities in the attached text are used to perform entity insertion according to the corresponding category. The blue words represent non-entities from the contextual sentences, and the red words represent entity placeholders and entities. (Color figure online)

2 Related Work

General Video Description. Early video descriptions [7, 8] were based on templates or rules to generate fixed-form sentences after detecting the subject, predicate, and object. With the development of deep learning, some deep learning models are applied to video descriptions. The S2VT [9] model used the CNN [10] to extract video frame features, and then used LSTM [11] for decoding. With the development of tasks such as object detection and vision classification, researchers began to use VGG [12], ResNet [13], etc.

to extract the appearance features of videos, and use 3D-CNN [14], I3D [15], etc. to extract motion features of videos and apply them to downstream tasks related to videos. [16, 17] comprehensively used 2D and 3D convolutional networks to extract static and dynamic features of videos to improve the accuracy of generated descriptions. Due to the lack of background knowledge, the sentences generated by these methods cannot contain named entities, which has great limitations.

News Video Description. News video description enriches the video description by introducing additional articles related to the videos. KaVD [3] extracted entities from the attached text, and used pointer network [18] and entity embedding to enable the model to output entities during decoding. S2VT-Pointer [4] proposed an end-to-end description model, which used all words in the attached text when decoding. However, these methods do not filter redundant contextual sentences, and both use a one-stage structure, which cannot fully mine text information. In news image description, [19] uses a template-based method to describe a news image, and [20] uses the pre-trained model Roberta [21] to encode text and use multilayer transformers as the decoder to achieve good results. Inspired by news image descriptions, this paper proposes a two-stage method based on template generation and entity insertion. First, a sentence ranker is designed to sort the contextual sentences, and then the contextual sentences and video are used to generate a description template containing entity placeholders. Finally, we use the entities of the sorted sentences for entity insertion by category.

3 Approach

Figure 2 shows the pipeline of our method. We first use the sentence ranker to rank the contextual sentences according to their relevance to the video and pick some sentences for subsequent description generation. Then a description template is generated by a template generation model based on a multimodal encoder and a transformer-based decoder. Based on the template, we use the entities of the sorted sentences for entity insertion by category to obtain the final video description.

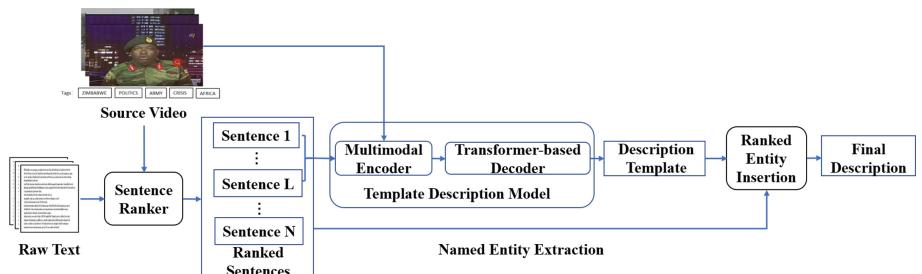


Fig. 2. The overall pipeline of our method.

3.1 Description Template Generation

Sentence Ranker. A video sample contains L_r frames $\{r_1, \dots, r_{L_r}\}$ and L_t video tags $\{t_1, \dots, t_{L_t}\}$. Sampling uniformly from the frames, we get the N RGB frames $\{r_1, \dots, r_N\}$ and N clips $\{c_1, \dots, c_N\}$, where each clip c_i consists of consecutive frames around each sampled frame r_i . We use 2D-CNN and 3D-CNN to extract the appearance feature $\{v_1^a, \dots, v_N^a\}$ and motion feature $\{v_1^m, \dots, v_N^m\}$ respectively. The two types of features are concatenated and fed through bidirectional LSTM from which we get its hidden states as the temporal visual feature X^V :

$$z_i = [v_i^a; v_i^m] \quad (1)$$

$$X^V = [X_1^V, \dots, X_N^V] = \text{LSTM}([z_1, \dots, z_N]) \quad (2)$$

where $X^V \in \mathbb{R}^{N \times d_v}$, d_v is the hidden size of LSTM.

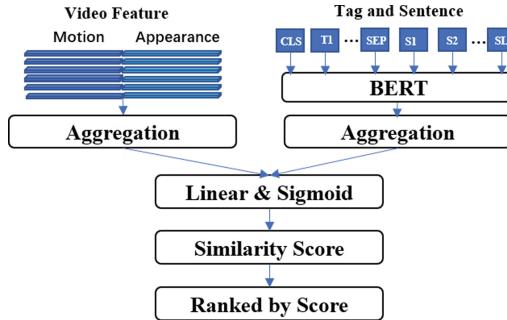


Fig. 3. The structure of the sentence ranker

Since a sample may have several articles, and there are a large number of sentences with low relevance to the video, it is necessary to filter the attached text. Inspired by [22], We design a sentence ranker based on the pre-trained model BERT. The BERT is a language model trained on a large corpus and has shown strong performance in many fields. Figure 3 shows the process of sentence ranker calculating the similarity between videos and sentences. We use BERT to encode video tags $\{t_1, \dots, t_{L_t}\}$ and a certain contextual sentence $\{s_1, \dots, s_{L_s}\}$ to get hidden text feature $\{x_1, \dots, x_{L_x}\}$:

$$\{x_1, \dots, x_{L_x}\} = \text{BERT}(\{[\text{CLS}], t_1, \dots, t_{L_t}, [\text{SEP}], s_1, \dots, s_{L_s}\}) \quad (3)$$

where $L_x = L_t + L_s + 2$, L_t is the number of video tags, L_s is the number of words in one sentence, $[\text{CLS}]$ and $[\text{SEP}]$ are the special tokens of BERT.

To calculate the similarity between contextual sentences and videos, inspired by [23], we use the attention mechanism to define the aggregation operation $\text{Agg}(\cdot; v)$ as follows:

$$\text{Agg}(x; w) = \sum_{i=1}^{L_x} \alpha_i x_i \quad (4)$$

$$\alpha_i = \text{softmax}(w^T x_i) \quad (5)$$

where $w \in R^d$ is a learnable parameter, can be regarded as a screening factor for the vector x , and weighted average to get the aggregated vector $\text{Agg}(x; w)$.

We aggregate the hidden text feature to get the contextual vector $e_x = \text{Agg}(x; w_x)$ and similarly, we aggregate the temporal visual features X^V to get the visual vector $e_v = \text{Agg}(v; w_v)$. Define the similarity of a video and a contextual sentence as follows:

$$\text{sim}(x, v) = \text{sigmod}(W_r[e_v, e_x] + b_r) \quad (6)$$

where w_x, w_v, W_r , and b_r are learnable parameters.

We use the ROUGE-L[24] of the contextual sentence x and the real description of the video v as the ground-truth, and train the model by minimizing the cross-entropy loss of $\text{sim}(x, v)$ and ground-truth. In testing, each contextual sentence in the attached text is sorted based on the predicted score to get a sorted list $[R_1, \dots, R_L]$, and then the first L' sentences are selected as the ranked sentences $[R_1, \dots, R_{L'}]$.

Multimodal Encoder. For video, X^V in Sentence Ranker is the visual feature after encoding. For text, we utilize the non-entities and entities of the picked sentences for template generation and entity insertion, respectively.

We extract the named entities for each contextual sentence and replace the entities with their corresponding entity categories to get ranked entities and the replaced sentence template. We also process the ground truth of the video description to get the reference template. To distinguish the replaced entity category from the original vocabulary, we add a specific “_” identifier.

For example, “At least 41 people are killed and hundreds injured after a strong earthquake struck off Aceh province on Indonesia’s Sumatra island”. After named entity extraction, we get the entity type of “At least 41” and “hundreds” are “Cardinal”, “Aceh” and “Indonesia” are “GPE” which means geopolitical entities, “Sumatra island” is “LOC” which means location. The sentence template is “_CARDINAL_ people are killed and _CARDINAL_ injured after a strong earthquake struck off _GPE_ province on _GPE_’s _LOC_”.

We extract named entities on the filtered sentences $[R_1 \dots R_{L'}]$ to get the sentence templates $[T_1, \dots, T_{L'}]$ and ranked entities $\{e_i, t_{ei}\}_{i=1}^{N_e}$, where $T_i = [s_{i1}, \dots, s_{iL_s}]$ is sentence template for sentence i , e_i is the i -th entity name, t_{ei} is the i -th entity category, N_e is the number of entities. The sentence templates are concatenated and embedded to get the contextual feature $X^C = [X_1^C, \dots, X_L^C]$, where $X^C \in R^{L \times d_x}$ (Fig. 4).

Transformer-Based Decoder. The decoder outputs the description word by word. At time step t , it takes as input: words generated in the previous step $S_{<t} = \{s_0, s_1, s_2, \dots, s_{t-1}\}$, the visual feature X^V , the contextual feature X^C . These inputs are fed through M DecoderLayers and generate hidden variables $H_t^i = \{z_1^i, \dots, z_t^i\}$, where $i \in \{0, 1, \dots, M\}$. H_t^0 is the embedding of the generated words $S_{<t}$. The process can be formulated as:

$$H_t^l = \text{DecoderLayer}_l(H_t^{l-1}, X^V, X^C) \quad (7)$$

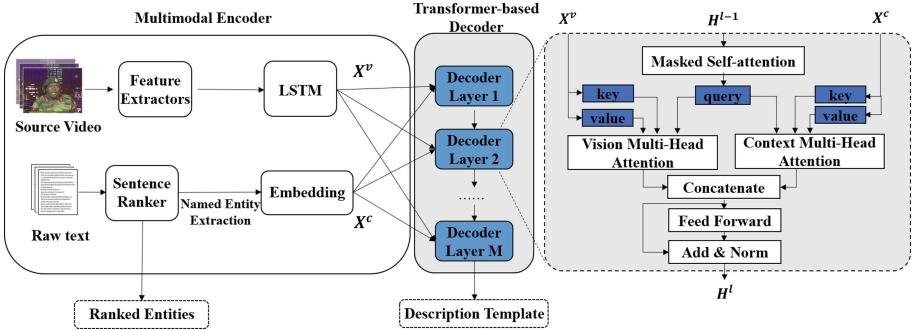


Fig. 4. Overview of our template generation model where we combine the visual and textual features to get the description templates.

where $l \in [1, M]$. Perform a linear projection and softmax operation on the t -th element of H_t^M to get a probability over words in the dictionary from which we can get the generated word \tilde{s}_t at time step t . The decoding process can be viewed as combining visual features and contextual features to obtain the next generated word under the guidance of the partially generated word:

$$\tilde{s}_t \leftarrow \text{Decoder}(S_{<t}, X^V, X^C) \quad (8)$$

DecoderLayer is based on the attention mechanism [6]. We first introduce scaled dot-product attention, the basic component of the transformer. For a set of queries Q , keys K , and values V , the value vectors are weighted and summed according to the similarity between the query and key vectors:

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (9)$$

The multi-head attention consists of H parallel scaled dot-product attention layers:

$$MA(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (10)$$

$$\text{head}_i = A\left(W_i^Q Q, W_i^K K, W_i^V V\right) \quad (11)$$

where the projections $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$, and $W^O \in R^{hd_v \times d_{model}}$ are learnable parameters.

For the l -th DecoderLayer, use the self-attention mechanism to get \tilde{H}_t^l :

$$\tilde{H}_t^l = MA\left(H_t^{l-1}, H_t^{l-1}, H_t^{l-1}\right) \quad (12)$$

After that, the multi-head attention mechanism is applied to the visual and contextual features respectively:

$$\tilde{x}_t^{vl} = MA\left(\tilde{H}_t^l, X^V, X^V\right) \quad (13)$$

$$\tilde{X}_t^{Cl} = MA\left(\tilde{H}_t^l, X^C, X^C\right) \quad (14)$$

After concatenation and linear projection, we get H_t^l which is used as input for the next DecoderLayer.

$$\tilde{X}_t^l = \begin{bmatrix} \tilde{X}_t^{Vl}; \tilde{X}_t^{Cl} \end{bmatrix} \quad (15)$$

$$\tilde{X}_t^{l'} = W_l \tilde{X}_t^l + b_l \quad (16)$$

$$\widetilde{X}_t^{l''} = ReLU\left(W'_l \tilde{X}_t^{l'} + b'_l\right) \quad (17)$$

$$H_t^l = LayerNorm\left(\widetilde{X}_t^{l'} + W''_l \widetilde{X}_t^{l''} + b''_l\right) \quad (18)$$

Training. Given a video V, the attached text C, and a ground-truth description S, we use Cross-entropy loss as the objective function for our training, as follows:

$$L = \sum_{(V,C,S) \in D} \sum_t (-\log p(s_t | V, C, s_0, s_1, \dots, s_{t-1})) \quad (19)$$

3.2 Named Entity Insertion

After generating the description template, we insert entities on the template according to the entity's category. The key problem is how to filter entity information from text, we design four methods to insert entity.

No Insertion. No entity is inserted, and the corresponding entity category is used to refer to the entity. This is the baseline of entity insertion.

Random Insertion. For each entity placeholder of the template, an entity of the corresponding category is randomly selected from the attached text.

Glove [25] Insertion. We use the average Glove embedding of words in the sentence as sentence embedding. Contextual sentences are ranked according to the cosine similarity between the description template and the contextual sentences, and named entities are inserted according to the ranking.

Ranked Entity Insertion. From Sentence Ranker we can get the ranked entities $\{e_i, te_i\}_{i=1}^{N_e}$, then insert entities based on entity category and ranking. Figure 5 is an example. There are placeholders for “ORG” and “PERSON” in the description template, and the corresponding entity “Zimbabwe Defense Forces” and “S.B. Moyo” are selected from the ranking entities to get the final description.

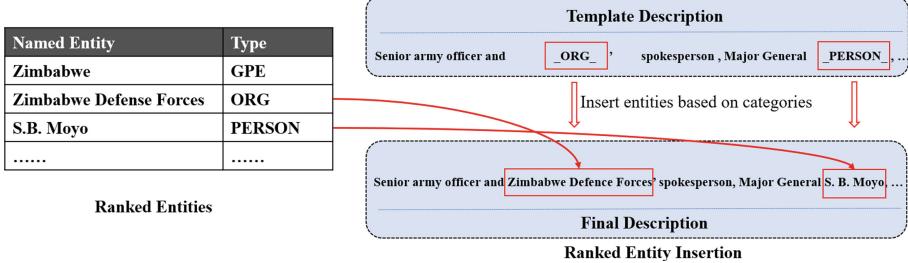


Fig. 5. An example of entity insertion.

4 Experiments

4.1 Dataset

News Video Dataset [3] is the most used dataset in the news video description. Each sample of the dataset contains a video, several video tags, and several attached articles which are retrieved from the video metadata tags. The videos are from October 2015 to November 2017 and cover a variety of topics such as protests, attacks, natural disasters, trials, and political movements. We downloaded 2809 available samples based on the link provided by the author. We use the same data split as other methods, randomly selecting 400 videos for testing, 80 for validation, and 2403 for training.

4.2 Training Details

For each video, $N = 30$ frames and clips are uniformly sampled from it. We use the I3D pre-trained on Kinetics 400 [15] to get the 1024-dimensional motion feature and use the Resnet50 pre-trained on 1k ImageNet [26] to get the 2048-dimensional appearance feature. Set the hidden size of LSTM to 1024.

In the sentence ranker, each pair of a contextual sentence and video is regarded as a training sample. The similarity between the real description and the video is set to 1 and added to the training set. The training set has a total of 323,840 samples. The bert-large-cased is used for text feature extraction. The batch size is 320, and Adam [27] is used for optimization, the learning rate is $1e-4$, $\beta = (0.9, 0.999)$. We apply L2 regularization to all network weights with a weight decay of $1e-5$. Fine-tune the model for 3 epochs. After sorting according to the similarity, we select the top 20 sentences to concatenate. The maximum length after concatenating is set to 300.

We use SpaCy [28] for named entity extraction in template generation. The number of heads in the transformer-based decoder is 8, the batch size is 16, the number of decoders is 5, and Adam is also used for optimization, the learning rate is $1e-4$, $\beta = (0.9, 0.999)$, the generated description length is 40. At inference time, we use greed search for description generation.

4.3 Evaluations Metrics

For Sentence Ranker, we choose Hit@N and Mean Rank to measure its performance. Hit@N means the percentage of real descriptions appearing in the top N after sorting. Mean Rank means the average of the real description rankings.

For the generated descriptions, we use BLEU [29], METEOR [30], ROUGE-L, and CIDEr [31] for evaluation which are commonly used in machine translation and video description. Besides, we also compare the F1 scores of the named entities and the F1 scores of “PERSON” which refers to the entity’s category identified as “PERSON” in the generated description. We compare entities with the same characters in the generated description and the real description.

4.4 Results and Analysis

Sentence Ranker. Table 1 shows the results of Sentence Ranker in the test set. The probability of the real description ranking first after sorting is 67.75%, and the average ranking is 3.225. The results show that Sentence Ranker can effectively identify sentences related to videos.

Table 1. Performance evaluation of Sentence Ranker

Model	Hit@1	Hit@5	Hit@10	Mean rank
Sentence ranker	67.75	87.50	94.25	3.225

Description Template Generation. The results of the generated description template are shown in Table 2. To prove the effectiveness of each module in the model, we also design ablation experiments to compare the metrics when only the context input, only the video input, and both context and video input. It can be seen that the model with only text input is better than only video input. When the context and video are input together, the results are the best which shows that our model can effectively mine information from contextual sentences and videos for template generation.

Table 2. Performance evaluation of description template generation for different inputs

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDER
Only context	39.59	13.34	20.41	33.01	12.78
Only video	37.84	12.00	18.80	30.40	9.30
Our model	39.84	13.77	20.81	33.13	13.06

Named Entity Insertion. After entity insertion, our model using different insertion methods is compared with two state-of-the-art models. Table 3 shows the results.

When no entity is inserted, it has poor description quality, BLEU-4, entity F1, and Person F1 are all 0. This is to be expected since a lot of information is lost in this way. Random insertion generates more accurate descriptions due to the introduction of entities compared to no insertion. Compared with random insertion, the Glove insertion has a great improvement in the CIDEr and F1 scores of entities and is close to KaVD and S2VT-Pointer in METEOR, ROUGE-L, and F1 scores of entities.

Table 3. Performance evaluation of entity insertion

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	Entity F1	Person F1
KaVD	×	10.2	18.9	×	22.1	×
S2VT-pointer	×	10.8	18.6	25.7	×	×
No insertion	0	5.69	12.59	0.02	0	0
Random insertion	1.74	8.11	16.04	10.71	11.03	13.71
Glove insertion	2.11	9.29	17.42	16.66	22.12	15.90
Ranked entity insertion	3.47	11.17	19.56	24.61	25.35	28.63

The ranked entity insertion achieves the best results and can generate more accurate descriptions compared to other inserting methods. It outperforms the Glove insertion by large margins of 7.95% and 12.73% in CIDEr and F1 scores of people’s names, respectively. Compared to state-of-the-art methods, it has the best scores in all metrics except CIDEr which is slightly lower than S2VT-Pointer. It outperforms the KaVD by margins of 0.97%, 0.67%, and 3.25% in METEOR, ROUGE-L, and F1 scores of people’s names, respectively. It outperforms the S2VT-Pointer by margins of 0.37%, and 0.66% in METEOR and ROUGE-L, respectively. Ranked entity insertion shows powerful results, especially on the F1 scores of entities and people’s names.

Figure 6 shows an example from our test set. We see that the video focuses on the “protesters gather in Barcelona to demonstrate”, our model can generate the corresponding description template “_ gather in _gpe_ to protest against”, and also contains other entity placeholders such as “_cardinal_”. Then entities of the corresponding category from the sorted sentences are selected for insertion. It can be seen that the final description correctly generates entities such as “hundreds”, “Catalan”, “Barcelona”, etc. The result shows that our model can generate accurate description templates and extract relevant entities from the attached text according to the templates.

	
Ranked Sentences	<p>1. People attend a protest in Barcelona on Monday, October 2, a day after hundreds were injured in a police crackdown during the banned referendum.</p> <p>2. People wait at the doors of a school in Barcelona to start voting during the Catalan independence referendum.</p> <p>3. Thousands of people gather in Barcelona to rally for unity in Spain on October 8.</p> <p>.....</p>
Reference description	Hundreds of protesters gather in Barcelona to demonstrate a day after hundreds were injured in a police crackdown during Catalonia's banned independence referendum
Reference template	<u>_cardinal_</u> of protesters gather in <u>_gpe_</u> to demonstrate <u>_date_</u> after <u>_cardinal_</u> were injured in a police crackdown during <u>_gpe_</u> 's banned independence referendum
Description template	<u>_cardinal_</u> of <u>_norp_</u> gather in <u>_gpe_</u> to protest against president <u>_person_</u> in <u>_gpe_</u> to protest against <u>_gpe_</u> 's main opposition leader <u>_person_</u> as he was shot down by the government
Final description	hundreds of Catalan gather in Barcelona to protest against president 01:05 in Spain to protest against Catalonia's main opposition leader Carles Puigdemont , as he was shot down by the government

Fig. 6. Sample prediction on news video dataset. Different colors: different entity categories. Underlined words: correct entities or events. (Color figure online)

5 Conclusion

In this paper, we propose a two-stage news video description model based on template generation and entity insertion. This structure allows the model to focus on the event in the template generation, and focus on entities such as event participants in the entity insertion. The results show that our model exhibits strong performance on various metrics of video description and entity accuracy. The accuracy of the generated description is limited by the attached text. In future work, we will explore how to use more multi-source context to improve the quality of the generated description.

Acknowledgments. This research was funded by the National Key Research and Development Program of China (No. 2019YFB2101600).

References

1. Zheng, Q., Wang, C., Tao, D.: Syntax-aware action targeting for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13096–13105 (2020)
2. Pan, B., Cai, H., Huang, D.A., et al.: Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10870–10879 (2020)

3. Whitehead, S., Ji, H., Bansal, M., et al.: Incorporating background knowledge into video description generation. In: 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 3992–4001. Association for Computational Linguistics (2018)
4. Rimle, P., Dogan-Schönberger, P., Gross, M.: Enriching video captions with contextual text. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 5474–5481. IEEE (2021)
5. Devlin, J., Chang, M.W., Lee, K., et al.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
6. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
7. Nagel, H.H.: A vision of ‘vision and language’ comprises action: an example from road traffic. In: Artif. Intell. Rev. **8**(2), 189–214 (1994)
8. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. In: Int. J. Comput. Vision **50**(2), 171–184 (2002)
9. Venugopalan, S., Rohrbach, M., Donahue, J., et al.: Sequence to sequence -- video to text. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE (2016)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, p. 25 (2012)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. 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)
13. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Ji, S., Xu, W., Yang, M., et al.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
15. Carreira, J., Zisserman, A.: 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)
16. Li, X., Zhao, B., Lu, X.: MAM-RNN: multi-level attention model based RNN for video captioning. In: IJCAI, pp. 2208–2214 (2017)
17. Pei, W., Zhang, J., Wang, X., et al.: Memory-attended recurrent network for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8347–8356 (2019)
18. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. arXiv preprint [arXiv:1704.04368](https://arxiv.org/abs/1704.04368) (2017)
19. Biten, A.F., Gomez, L., Rusinol, M., et al.: Good news, everyone! Context driven entity-aware captioning for news images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12466–12475 (2019)
20. Tran, A., Mathews, A., Xie, L.: Transform and tell: entity-aware news image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13035–13045 (2020)
21. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: a robustly optimized Bert pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019)
22. Liu, Y., Lapata, M.: Hierarchical transformers for multi-document summarization. In: arXiv preprint [arXiv:1905.13164](https://arxiv.org/abs/1905.13164) (2019)
23. Zhang, Z., Qi, Z., Yuan, C., et al.: Open-book video captioning with retrieve-copy-generate network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9837–9846 (2021)

24. Rouge, L.C.Y.: A package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization of ACL, Spain (2004)
25. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
26. Deng, J., Dong, W., Socher, R., et al.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
27. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
28. Honnibal, M., Montani, I.: Natural language understanding with Bloom embeddings, convolutional neural networks, and incremental parsing. Unpublished software application (2017). <https://spacy.io>
29. Papineni, K., Roukos, S., Ward, T., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
30. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
31. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)



Local Feature for Visible-Thermal PReID Based on Transformer

Quanyi Pu¹(✉), Changan Yuan^{2,3}, Hongjie Wu⁴, and Xingming Zhao⁵

¹ Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
2030772@tongji.edu.cn

² Guangxi Academy of Science, Nanning 530007, China

³ Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision, Guangxi Academy Sciences, Nanning, China

⁴ School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

⁵ Institute of Science and Technology for Brain Inspired Intelligence (ISTBI), Fudan University, Shanghai 200433, China

Abstract. Person re-identification based on infrared image and RGB image is a cross-modality pedestrian recognition, which is a challenging task. The traditional goal of person re-identification is to find a given person's image from an image database, often from a single modality database. In real applications, there are often multiple modalities of data. Traditional single modality tasks have limitations. Cross-modality person re-identification needs to extract features from RGB and infrared images. In our work, we take advantage of both global and local features. First, we use a dual-path ViT structure to extract features from RGB images and infrared images, respectively. Secondly, we cut the local features in the spatial direction and input the shared ViT layer to learn the local features. The loss function consists of Identity loss, Triplet loss, and Center loss. The model can capture shared features between modality and improve cross-modality similarity. Finally, we performed experiments on two datasets, SYSU-MM01 and RegDB, and compared them with other methods in recent studies.

Keywords: Deep learning · Person re-identification · Local feature · Cross-modality · Vision transformer · Dual-path structure

1 Introduction

Person re-identification is a technology that uses computer vision algorithm to judge whether there are specific pedestrians in images or video sequences. It is usually considered as a sub problem of image retrieval. Using recognition technology, specific pedestrian targets can be retrieved from cross device images and videos, and can make up for the limitation of viewing angle under the current fixed camera [1]. Combined with pedestrian detection or tracking technology, this technology is often used in intelligent security, intelligent video surveillance, intelligent retrieval and other fields [2].

The research focuses on the current situation of PReID and some deep learning network structures [3–11, 35–38]. PReID task is similar to other computer vision tasks, which has some problems, such as image occlusion, light spot, illumination brightness and so on. At the same time, different from other computer vision tasks, PReID tasks also have some unique problems, such as the change of pedestrian posture, the change of camera shooting angle and low camera resolution.

Traditionally PReID is focused on single modality, and most of them are applied to scenes with sufficient light source. With the continuous improvement of video security monitoring requirements, in order to overcome the disadvantage that visible cameras cannot be used all day, cameras that can switch infrared mode are becoming popular. Therefore, in real scene applications, infrared mode cameras, depth cameras and pedestrian images captured and described by eyewitness statements are very common. In the past, the data sets used for training and testing are often single-modality RGB images. For the problem of different lighting conditions in day and night, cross-modal PReID is proposed. PReID across visible and infrared modes is one of the urgent problems to be solved. Cross modality PReID mainly studies the problem of trying to retrieve and match the image belonging to the same individual in the image database under the two modalities in the visible image or infrared image of a given individual.

For the problem of cross-modality identification, some PReID methods have been proposed. A generative adversarial network (GAN) training method proposed by Dai et al. named cmGAN [12, 39, 40], the triplet loss and cross entropy loss are combined. This method tries to learn the distinctive feature representation of different pattern matching. Wu et al. contributed a cross-modality PReID dataset called SYSU-MM01 [13, 41–43] and proposed a deep zero-padding model to address the cross-modality issue. Wang et al. introduced the AlignGAN method (Wang et al., 2019a) [14] to reduce intra modal changes by combining pixel alignment and feature alignment.

However, many current methods only focus on the global features and ignore the local feature representation. Due to the influence of the characteristics of pedestrian recognition task itself, the changeable environment, including illumination, camera angle of view, pedestrian definition, posture, gait, etc., will have a great impact on the global characteristics [44–47]. The local characteristics of the same pedestrian are usually unchanged, such as hair style, hat or shoe style and so on. Therefore, paying attention to local features cannot be affected by cross modal images. In our work, in order to better extract local and global features and break through the limitations of CNN, we use the dual-path model based on VIT. The model contains two separate branches, one is visible flow and the other is infrared flow, which extract the features of each mode respectively. Compared with CNN, transformer model shows advantages in single-modality PReID (He et al. 2021) [15]. It obtains the global receptive field through the self-attention modules, and has complete spatial features without the need for pooling layers. In the VIT model, the similarity between the input image and the feature map of the last layer is very high [48, 49]. This shows that VIT not only propagates the feature information, but also retains the location information. In addition, in our work, we split the feature map learned by VIT and jointly learn local features with global features. Finally, we conducted experiments on two public datasets SYSU-MM01 and RegDB, and achieved good performance.

2 Related Work

Visible-Infrared Person Re-identification. Due to the huge gap between RGB domain and infrared domain, a lot of work has been proposed to solve the problem of cross modal matching. Wu et al. [16] first proposed the definition of cross modality PReID in the field of PReID, analyzed three network architectures, proposed a data preprocessing method of deep zeroing, and compared and evaluated the performance of these four networks. Ye et al. [17] proposed a hierarchical cross modal matching modality, which is realized by jointly optimizing modality specificity and modality sharing matrix. Its framework is divided into two parts: representation learning and measurement learning. The former constructs a two stream network to learn the features of image inputs belonging to two modalities, and then combines the feature loss and contrast loss to learn the similarity. Dai et al. [18] in order to solve the problem of insufficient identification information, using the idea of confrontation training of Gan generator and discriminator, proposed a cross pattern generation confrontation network, which is divided into two parts: generator and discriminator. Then, in order to solve the two problems of differences between and within modes, Liu et al. [19] proposed an enhanced the discriminative feature learning (EDFL), which integrates the middle layer features by using jump connection to enhance the robustness of the features. The method based on graph convolution network has also achieved good results in the problem of cross modality PReID. Zhang et al. [20] proposed a new graph convolution model PGCN, which can learn the local inter relationship and local intra relationship of feature representation at the same time.

Vision Transformer. In order to solve some problems in the field of natural language processing, researchers proposed Transfomer (Vaswani et al. 2017) [21], which and its variants can deal with many problems in this field and dominate this field for a period of time. In recent years, with the advantages of self attention mechanism, many researchers have applied Transformer to the fields of computer vision and speech processing. Compared with convolution neural network, it has unique advantages. Such as image classification based on visual transformation model (Dosovitskiy et al. 2021) [22].

3 Proposed Method

In this section, we will introduce the framework model of our proposed cross modal preid task. As shown in Fig. 1, the model consists of the following three parts: (1) Dual-path VIT structure, (2) Local feature processing module, (3) Identity Loss, Triplet Loss and Center Loss.

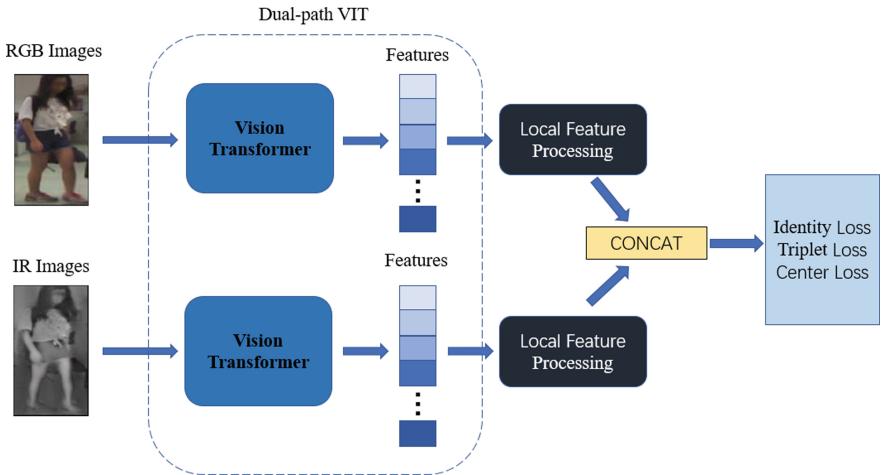


Fig. 1. This is the overall structure of our model. We selected VIT as the backbone. Two independent VIT modules of dual-path extract the features of RGB and IR images respectively. In the local feature processing module, the local feature will be segmented and combined, and will be combined with the global feature through a separate transformer layer. The proposed model is trained by Identity Loss, Triplet Loss and Center Loss.

3.1 Dual-Path VIT Structure

In order to extract the cross modality character features of infrared and visible images, we use a dual-path structure. Because VIT has achieved excellent performance in many visual tasks recently, we selected two independent VIT networks as the backbone of each path to extract the character features of each modality image respectively. Give a figure picture $x \in R^{H \times C \times W}$, where H, W and C are height, width and channel dimensions. Use a sliding window, set the stride as S, and the window is a square with side length P, then the number of patches $N = (H + S - P/S) \times (W + S - P/S)$. In our experiment, taking stride S as 1/2 of P and segmenting patches based on overlap can effectively enhance the connection of each patches and alleviate the lack of local relationship.

The independent VIT model is helpful to extract the character features of specific modality and alleviate the problem of cross modality change. Global features can provide global information for each person image, but also take into account the loss of local information. After embedding the position coding P_E , the obtained feature map will be cut into multiple horizontal stripes in the next step and trained in combination with the global features.

3.2 Local Feature Processing Module

In order to obtain the local feature representation of each image, we cut the global feature graph. A person's image can be sliced into multiple horizontal stripes, and each stripe focuses on a specific local information. The sliced feature map is spliced with the token, and each local feature map is input into the BN layer of weight initialization to speed up the training and convergence of the network and prevent over fitting. When the number

of sliced blocks is too large, the characteristic information of each block will be lost. When the number of segmentation blocks is too small, the extraction effect of local significant information is not obvious. In our experiment, we slice the feature map into 8 pieces (Fig. 2).

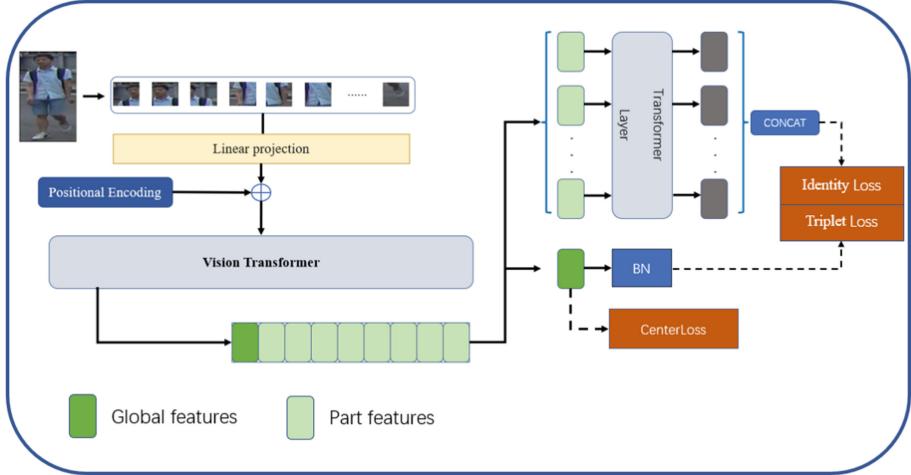


Fig. 2. This is the detailed structure diagram of our model. After a picture is sliced, it is input into the VIT module with position code to obtain the global feature diagram. Then cut the feature map horizontally into several feature maps of the same size, and input an independent Transformer Layer. The obtained local features are combined with the global features after BN layer through a variety of loss functions for joint training.

The local feature map obtained after that will be input into the second independent transformer layer. The basic unit of each transformer layer consists of two parts: the block module in the VIT model and an independent linear layer. A basic unit will be repeated four times, enough to extract the local features of each slice.

At the same time, in order to obtain the global feature information, the global feature map will also be jointly trained with local features after going through BN layer.

3.3 Loss Functions

We use three loss functions in total. Because the training process of PReID can be regarded as an image classification problem, and different pictures of the same pedestrian can be regarded as a category, ID loss can be used.

$$L_{ID} = -\frac{1}{n} \sum_{i=1}^n \log(p(y_i|x_i)) \quad (1)$$

where n is the number of samples trained in each batch, and $p(y_i|x_i)$ is the input image x_i and its category label y_i . After softmax classification, x_i is recognized as the prediction probability of y_i class, which is expressed by $p(x_i|y_i)$.

At the same time, the training process of PReID model can also be regarded as a retrieval and sorting problem. The basic idea is that through the predefined margin, the distance between positive pairs should be less than that between negative pairs. Triple loss includes an anchor sample, a positive sample and a negative sample.

$$L_{TRI}(i, j, k) = \max(\rho + d_{ij} - d_{ik}, 0) \quad (2)$$

where d_{ij} , d_{ik} represents the Euclidean distance between anchor sample and positive sample and negative sample respectively. ρ is a margin parameter and a hyperparameter.

Consistent with the idea of metric learning, we hope that similar samples are compact and different samples are scattered. Therefore, we use center loss to train the global features in the network. The depth features usually have strong discrimination, that is, strong inter class discrimination.

$$L_C = \frac{1}{2} \sum_{i=1}^m ||x_i - c_{y_i}||_2^2 \quad (3)$$

c_{y_i} is the category center corresponding to each sample in each batch.

4 Experiments

4.1 Dataset Description

We used two common data sets in the experiment. SYSU-MM01 [13] is the largest data set for cross modality PReID. There are photos of six cameras in total, including 4 visible cameras and 2 infrared cameras. And three cameras shoot outdoors and three cameras shoot indoors. There are 491 identities in the data set, and each identity is captured by at least one visible camera and one infrared camera. There are 287628 RGB images and 15729 infrared images in total. There are 395 different identities on the training set of the whole data set, with a total of 22258 RGB images and 11909 infrared images. The test set has 96 different identities, with a total of 301 RGB images and 3803 infrared images.

RegDB is taken by two cameras [23]. There are 8240 images in the data set, including 412 identities. The training set and the test set each contain 206 identities. Each identity has 10 different RGB images and 10 different infrared images. Because these images are taken while people are moving, each person's 10 images differ in body posture, capture distance and lighting conditions. However, in the 10 images of the same person, the weather condition, viewing angle and shooting angle (front and rear viewing angle) of the camera are the same. Cumulative Matching Characteristics (CMC) and mean Average Precisions (mAP) are applied in all experiments (Fig. 3).

4.2 Implementation Details

All experiments were based on pytorch and two Titan XP GPUs. We use VIT-b16 pre training network as the backbone network. All images are resized to 224×224 . Random erase and horizontal random flip methods are used for data expansion. The initial learning rate was set to 0.001, and the learning rate decreased by 0.1 in epochs 5 and 25. The optimizer uses adamw, and decay is set to 0.005. The batch size is set to 4. The stride is set to 8 and the sliced score is set to 8.

**Fig. 3.** Examples from the SYSU-MM01 and RegDB.

4.3 Comparison with State-of-the-Art Methods

We compare the proposed model with other state-of-the-art methods, including TONE [24], TONE + HCML [24], BCTR [25], BDTR [25], cmGAN [12], D²RL [26], AlignGAN [14], CMGN [27], JSIA-ReID [28], XIV [29], MACE [30], DFE [31].

Table 1. Comparisons on SYSU-MM01 under all-search single-shot mode

Methods	Publication	Rank-1	Rank-10	Rank-20	mAP
TONE [24]	AAAI 2018	12.52	50.72	68.60	14.42
TONE + HCML [24]	AAAI 2018	14.32	53.16	69.17	16.16
BCTR [25]	IJCAI 2018	16.12	54.90	71.47	19.15
BDTR [25]	IJCAI 2018	17.01	55.43	71.96	19.66
cmGAN [12]	IJCAI 2018	26.97	67.51	80.56	27.80
D ² RL [26]	CVPR 2019	28.90	70.60	82.40	29.20
AlignGAN [14]	ICCV 2019	42.40	85.00	93.70	40.70
CMGN [27]	Neurocom2020	27.21	68.19	81.76	27.91
JSIA-ReID [28]	AAAI 2020	38.10	80.70	89.90	36.90
XIV [29]	AAAI 2020	49.92	89.79	95.96	50.73
MACE [30]	TIP 2020	51.64	87.25	94.44	50.11
DFE [31]	ACMMM 2019	48.71	88.86	95.27	48.59
Ours	—	52.91	89.80	96.32	52.99

We evaluated our proposed model on the SYSU-MM01 dataset. Table 1 shows the comparison results of Rank-n ($n = 1, 10, 20$) accuracy and mAP between our model and other methods on SYSU-MM01. It can be seen that the proposed model can

achieve Rank-1 accuracy of 52.91%, rank-10 accuracy of 89.80%, rank-20 accuracy of 96.32% and map 52.99% % performance. In terms of Rank-1 accuracy, we improved the performance of baseline from 44.67% to 52.91%.

Table 2. Comparisons on RegDB with Visible to Thermal mode

Methods	Publication	Rank-1	Rank-10	mAP
TONE [24]	AAAI 2018	16.87	34.03	14.92
TONE + HCML [24]	AAAI 2018	24.44	47.53	20.80
BCTR [25]	IJCAI 2018	32.67	57.64	30.99
BDTR [25]	IJCAI 2018	33.47	58.42	31.83
D ² RL [26]	CVPR 2019	43.40	66.10	44.10
AlignGAN [14]	ICCV 2019	57.90	—	53.60
CMGN [27]	Neurocom2020	35.13	61.07	32.14
JSIA-ReID [28]	AAAI 2020	48.50	—	49.30
XIV [29]	AAAI 2020	62.21	83.13	60.18
LBA [32]	ICCV 2021	74.17	67.64	—
VSD [33]	CVPR 2021	73.20	71.60	—
NFS [34]	CVPR 2021	80.54	91.96	72.10
Ours	—	81.60	95.34	74.35

We further evaluated our model on RegDB dataset. As shown in Table 2, we achieved an accuracy of 81.60% for Rank-1, 95.34% for Rank-10, and 74.35% for mAP performance. In terms of Rank-1 accuracy, we improved the performance of baseline from 71.03% to 81.60%, which further proves the effectiveness of the proposed model.

5 Conclusion

In this work, we use Vision Transformer as the baseline. Dual-path VIT is used to extract image features, one is visible flow, the other is infrared flow, and overlapping cutting is used to enhance the relationship between local areas. In order to obtain the local salient feature information, the extracted feature map is sliced, and the local feature and global feature are used for joint training. Finally, we use Center loss alone for global features and ID loss and Triplet loss for joint supervision of local features and global features. Our proposed method has achieved a great performance on SYSU-MM01 and RegDB datasets.

Acknowledgements. This work was supported by the grant of National Key R&D Program of China (No. 2018AAA0100100 & 2018YFA0902600) and partly supported by National Natural Science Foundation of China (Grant nos. 61732012, 62002266, 61932008, and 62073231), and

Introduction Plan of High-end Foreign Experts (Grant no. G2021033002L) and, respectively, supported by the Key Project of Science and Technology of Guangxi (Grant no. 2021AB20147), Guangxi Natural Science Foundation (Grant nos. 2021JJA170204 & 2021JJA170199) and Guangxi Science and Technology Base and Talents Special Project (Grant nos. 2021AC19354 & 2021AC19394).

References

1. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv preprint [arXiv:1610.02984](https://arxiv.org/abs/1610.02984) (2016)
2. Ye, M.: Deep learning for person re-identification: a survey and outlook (2020)
3. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: 2006 IEEE Computer Society Conference on IEEE Computer Vision and Pattern Recognition, pp. 1528–1535 (2006)
4. Bazzani, L., Cristani, M., Perina, A., et al.: Multiple-shot person re-identification by HPE signature. In: 2010 20th International Conference on Pattern Recognition, pp. 1413–1416. IEEE (2010)
5. Farenzena, M., Bazzani, L., Perina, A., et al.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2360–2367. IEEE (2010)
6. Wu, Y., Qin, X., Pan, Y., et al.: Convolution neural network based transfer learning for classification of floers. In: 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP), pp. 562–566. IEEE (2018)
7. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Pp. 4353–4361 (2015)
8. Yuan, C., Wu, Y., Qin, X., et al.: An effective image classification method for shallow densely connected convolution networks through squeezing and splitting techniques. *Appl. Intell.* **49**(10), 3570–3586 (2019)
9. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: what features are important? In: Fusillo, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_39
10. Wu, Y., Zhang, K., Wu, D., Wu, Y., et al.: Person reidentification by multiscale feature representation learning with random batch feature mask. *IEEE Trans. Cogn. Dev. Syst.* **13**(4), 865–874 (2021)
11. Wu, Y., et al.: Position Attention-Guided Learning for Infrared-Visible Person Re-identification. In: Huang, De-Shuang., Bevilacqua, Vitoantonio, Hussain, Abir (eds.) ICIC 2020. LNCS, vol. 12463, pp. 387–397. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60799-9_34
12. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross modality person re-identification with generative adversarial training. In: IJCAI, pp. 677–683 (2018)
13. Wu, A.: RGB infrared cross modality person re identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
14. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: RGB-infrared cross-modality person reidentification via joint pixel and feature alignment. In: ICCV, pp. 3622–3631 (2019)
15. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: transformer-based object reidentification. CoRR, abs/2102.04378 (2021)

16. Wu, A., Zheng, W.S., Yu, H.X., et al.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
17. Ye, M., Lan, X., Li, J., et al.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
18. Dai, P., Ji, R., Wang, H., et al.: Cross-modality person re-identification with generative adversarial training. In: IJCAI, vol. 1, p. 2 (2018)
19. Liu, H., Cheng, J., Wang, W., et al.: Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. Neurocomputing **398**, 11–19 (2020)
20. Zhang, Z., Zhang, H., Liu, S., et al.: Part-guided graph convolution networks for person re-identification. Pattern Recogn. **120**, 108155 (2021)
21. Vaswani, A., et al.: 2017. Attention is all you need. In NIPS, pp. 5998–6008 (2017)
22. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In ICLR (2021)
23. Dat, T.N., Hyung, G.H., Ki, W.K.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
24. Ye, M., Lan, X., Li, J.: Hierarchical discriminative learning for visible thermal person re-identification. In: Thirty Second AAAI Conference on Artificial Intelligence (2018)
25. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 1092–1099 (2018)
26. Wang, Z., Wang, Z., Zheng, Y.: Learning to reduce dual level discrepancy for infrared visible person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 618–626 (2019)
27. Jiang, J., et al.: A cross-modal multi-granularity attention network for RGB-IR person re-identification. Neurocomputing, p. 406 (2020)
28. Wang, G.-A., Zhang, T., Yang, Y.: Cross-modality paired images generation for RGB-infrared person re-identification. In: Thirty-Fourth AAAI Conference on Artificial Intelligence (2020)
29. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an X modality. In: Thirty-Fourth AAAI Conference on Artificial Intelligence (2020)
30. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Trans. Image Process. **29**, 9387–9399 (2020)
31. Hao, Y., Wang, N., Gao, X., Li, J., Wang, X.: Dual-alignment feature embedding for cross-modality person re-identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 57–65 (2019)
32. Tian, X., Zhang, Z., Lin, S., et al.: Farewell to mutual information: variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1522–1531 (2021)
33. Tian, X., et al.: Farewell to mutual information: variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1522–1531 (2021)
34. Chen, Y., et al.: Neural feature search for RGB-infrared person re-identification. arXiv preprint [arXiv:2104.02366](https://arxiv.org/abs/2104.02366) (2021)
35. Huang, J., Huang, D.S.: Deep reinforcement learning based trajectory pricing on ride-hailing platforms. ACM Trans. Intell. Syst. Technol. vol. 13, no. 3, Article 41 (2022)
36. Wu, Y., et al.: Person re-identification by multiscale feature representation learning with random batch feature mask. IEEE Trans. Cogn. Dev. Syst. **13**(4), 865–874 (2021)
37. Wu, D., Wang, C., Wu, Y., Wang, Q.-C., Huang, D.S.: Attention deep model with multi-scale deep supervision for person re-identification. IEEE Trans. Emerg. Top. Comput. Intell. **5**(1), 70–78 (2021)

38. Liang, X., Wu, D., Huang, D.S.: Image co-segmentation via locally biased discriminative clustering. *IEEE Trans. Knowl. Data Eng.* **31**(11), 2228–2233 (2019)
39. Wu, D., et al.: Deep learning based methods for person re-identification: a comprehensive review. *Neurocomputing* **337**, 354–371 (2019)
40. Wu, D., et al.: Random occlusion-recovery for person re-identification. *J. Imaging Sci. Technol.* **63**(3), 30405-1–30405-9(9) (2019)
41. Li, B., Fan, Z.T., Zhang, X.L., Huang, D.S.: Robust dimensionality reduction via feature space to feature space distance metric learning. *Neural Netw.* **112**(4), 1–14 (2019)
42. Wu, D., et al.: Omnidirectional feature learning for person re-identification. *IEEE Access* **7**, 28402–28411 (2019)
43. Wu, D., Zheng, S.-J., Yuan, C.-A., Huang, D.S.: A deep model with combined losses for person re-identification. *Cogn. Syst. Res.* **54**, 74–82 (2019)
44. Wu, D., Zheng, S.-J., Bao, W.-Z., Zhang, X.-P., Yuan, C.-A., Huang, D.S.: A novel deep model with multi-loss and efficient training for person re-identification. *Neurocomputing* **324**, 69–76 (2019)
45. Peng, C., Zou, L., Huang, D.S.: Discovery of relationships between long non-coding RNAs and genes in human diseases based on tensor completion. *IEEE Access* **6**, 59152–59162 (2018)
46. Yang, B., Bao, W., Huang, D.S., Chen, Y.: Inference of large-scale time-delayed gene regulatory network with parallel mapReduce cloud platform. *Sci. Rep.* **8**, 17787 (2018)
47. Shen, Z., Bao, W.-Z., Huang, D.S.: Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep.* **8**, 15270 (2018)
48. Liu, B., Weng, F., Huang, D.S., Chou, K.-C.: HSCVFNT: inference of time-delayed gene regulatory network based on complex-valued flexible neural tree model. *Int. J. Mol. Sci.* **19**(10), 3178 (2018)
49. Zhang, H., Zhu, L., Huang, D.S.: DiscMLA: an efficient discriminative motif learning algorithm over high-throughput datasets. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **15**(6), 1810–1820 (2018)



A Hardware Implementation Method of Radar Video Scanning Transformation Based on Dual FPGA

Naizhao Yu^(✉), Xiao Min^(✉), and Liang Zhao^(✉)

The 723th Research Institute of China Shipbuilding Industry Corporation,
Yangzhou 225001, China

laoyu110@163.com, ronaldom9@163.com, 490829963@qq.com

Abstract. A hardware implementation method of radar video scanning transformation based on dual FPGA is proposed. This method is a dual FPGA architecture. FPGA B is responsible for the realization of radar scanning transformation and afterglow control. After the radar video is sent to fpga A through SRIO, fpga B superimposes the radar video and DVI video. The method framework has the functions of bow/true north switching, range switching, self inspection, afterglow and local amplification. The overall structure is simple, the decoupled architecture has high reliability and strong expansibility.

Keywords: FPGA · Radar video scanning transformation · Afterglow

1 Background

The luminous color, luminous efficiency and afterglow time of the traditional fluorescent screen used in old-fashioned radar are related to fluorescent substances. The accumulation of afterglow can greatly improve the ability of operators to distinguish targets in noise or clutter. Modern display and control terminals need to save video data for a long time, which will occupy a lot of memory and affect the work efficiency of CPU. In view of the large amount of data and fast data update of radar video display, especially for the wake afterglow display function, the design based on hardware FPGA can release the CPU and improve the work efficiency of the CPU. At the same time, the parallel processing characteristics of FPGA are suitable for the characteristics of large amount of radar video display data and fast data update.

FPGA, Field Programmable Logic Gate Array. Using FPGA is a mainstream technology of digital signal processing. At present, FPGA is widely used in communication, digital control, medicine, audio and video processing, instrumentation, radar signal processing and other fields. FPGA has greatly promoted the development of integrated circuits and changed the traditional “bottom-up” design mode, making the digital system miniaturized, large capacity, fast speed, low power consumption and more flexible and convenient to use increasingly. When designing circuits, designers can follow “up-bottom” layer, do simulation verification by each layer, which improves the efficiency of design.

At present, the commonly used programmable logic devices mainly include simple logic array (PAL/GAL), complex programmable logic array (CPLD) and field programmable logic array (FPGA). Since FPGA has more triggers and more flexible wiring than CPLD, FPGA is more suitable for complex timing logic, which makes FPGA more widely used than CPLD.

FPGA consists of six basic parts: programmable input/output unit (I/O), basic programmable logic unit, embedded block ram, rich wiring resources, underlying embedded functional unit and embedded special hard core. In particular, the basic variable logic unit of FPGA is almost composed of look-up table (LUT) and output latch. For example, xc3s1000 of Xilinx spartan-iii series has 15360 LUTS. The amount of LUT measures the resource size of FPGA. In terms of common terminology, a three input LUT is written as 3-lut, and the same is true for LUTS of other sizes. The latch is used to register the output. Some FPGAs can provide all the outputs, while others choose whether to output registered signals or non registered signals through multiplexers. Registers can be used to delay data, establish finite state machines, and register data in pipelined systems.

The research work shows that the optimal size of LUT is to have 4 -6 inputs. Most of the early FPGA used 3-input LUT and 4-input LUT. However, due to the gradual reduction of the device size, the proportion of the total transmission delay caused by the wiring matrix has increased. This also makes the current devices have a large LUT. For example, Xilinx has a 6-input LUT. Another way to solve the above problem is to connect the two basic lookup tables and make them share some inputs. It is also common to combine multiple logical units into one function block or logic block. All logical units in the function block share common control signals, such as clock and clock enable signal. In the logic function block, the output signal in the same logic block can be directly used as the input of other logic units in the block. By connecting directly, the number of signals that need to be wired on the interconnection matrix can be reduced, so as to reduce the transmission delay on the critical path. Through a small number of additional dedicated multiplexers, adjacent LUTS can be connected, so as to synthesize more complex functions and establish larger LUTS.

The kintex 7 series 325t development board used by the author has 326080 logic units and uses 6-input LUT. In addition, Xilinx series development boards have rich IP cores (intellectual property cores) that can be used, such as commonly used operation cores such as multiplier core, divider core and FFT core, functional IP cores such as color system conversion core: ycbycr color system to RGB color system, storage cores such as block ram core, MIG core, clock frequency division IP core, etc.

2 System Composition and Work Flow

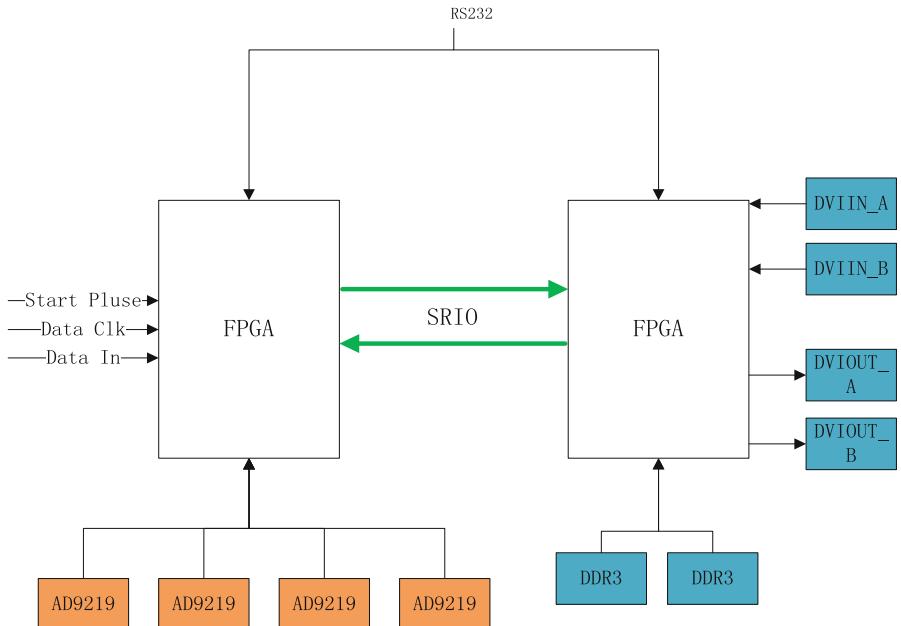


Fig. 1. Hardware system block diagram

As shown in Fig. 1, FPGA A is used for video display, FPGA is externally connected with two channels of DVI video, and two pieces of DDR3 are externally connected for video storage. Radar video transmitted by FPGA B is received through SRIO. FPGA B is used for radar video processing. The amplitude is calculated through AD sampling, and the azimuth code and range sweep pulse are calculated through the external analog radar video; Convert radar video to pixels using CORDIC. The calculated pixel address and brightness value are sent to FPGA A for display through SRIO.

DVI decoding chip use TFP401, DVI encoding chip use TFP410, DDR3 use MT41J256M16RE.

2.1 Design of FPGA A

Through SRIO, the radar video converted into pixels is received and stored in DDR3 with DVI input video to form a $1280 * 1024$ whole picture. After receiving the Uart command, the coordinates of the area to be amplified are obtained. After the area is amplified by the local amplification module, it is spliced with DVI video, written into DDR3 and then read out to obtain high-quality DVI video output (Fig. 2).

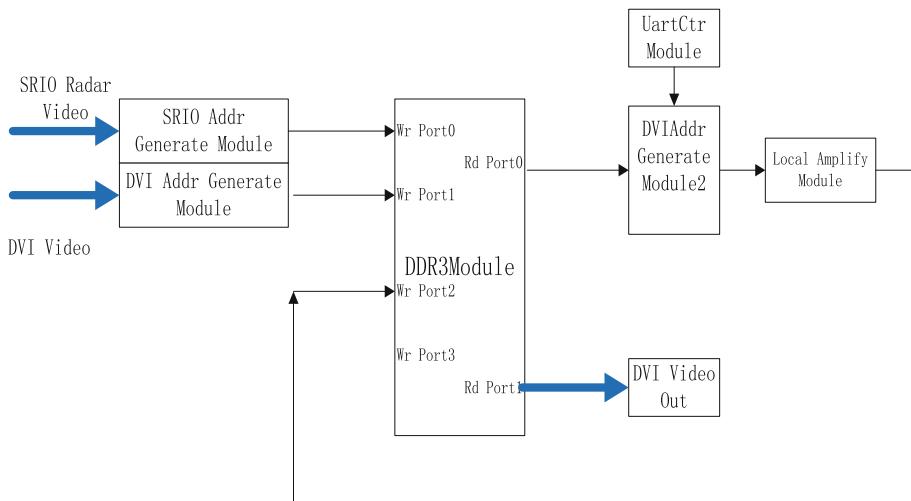


Fig. 2. Functional block diagram of FPGA A

2.2 Design of FPGA B

Through AD sampling module, azimuth code solution module and pixel calculation module, the input analog radar video is converted into the color value of pixels in the $1024 * 1024$ area on the left and sent to Bram afterglow control module (Fig. 3).

The VESA module is used to generate the DVI timing of $1280 * 1024 * 75$ Hz, and takes the field synchronization signal as the time reference. At the same time, the read enable of FIFO and blockram of afterglow control module is generated.

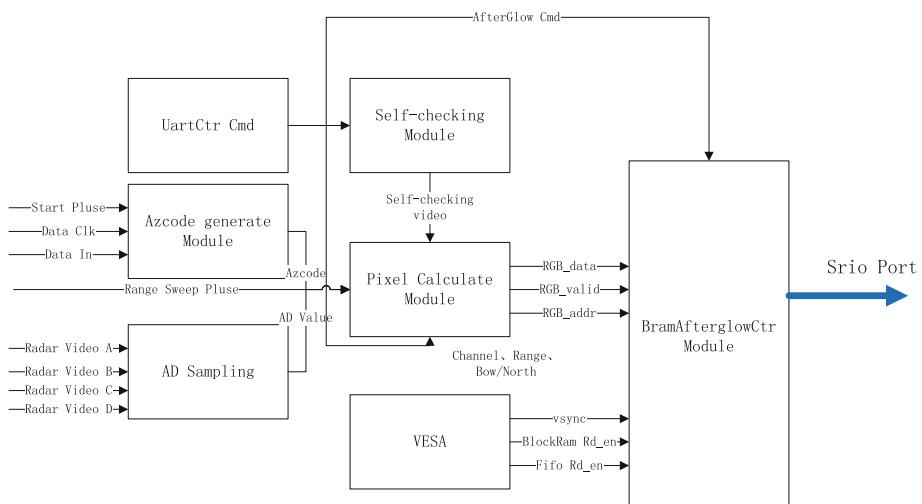


Fig. 3. Functional block diagram of FPGA B

The Uart command module receives and parses the serial command sent by the upper computer. Including afterglow gear, bow/due north switching, self inspection, range switching, local amplification, etc.

The self-checking module is used to generate simulated radar video.

Bram afterglow control module uses FIFO to cache 5M bandwidth radar video data and store it in blockram. According to the UART port message, when the time reaches the updated afterglow time threshold, attenuate the radar data in blockram to achieve the effect of radar video afterglow display. At the same time, the radar video is also sent to FPGA A through SRIO for radar video display.

3 Module Design

3.1 Conversion from Radar Video to Pixels

TIMING DIAGRAMS

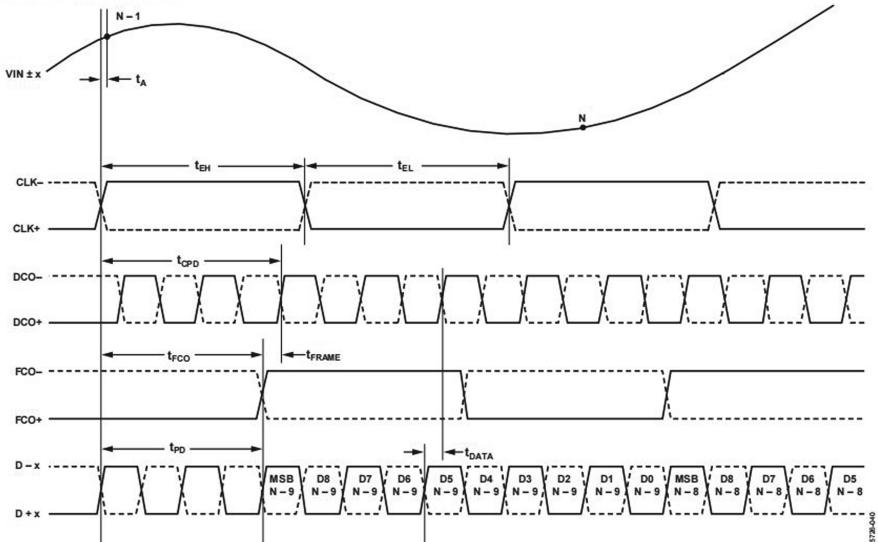


Fig. 4. AD9219 timing diagrams

- (1) The external start pulse, data input and data clock signal are input to the azimuth code solution module, and then the MMCM IPCore inside the FPGA outputs a 5M clock to the azimuth code solution module to obtain the azimuth code under the 5M bandwidth.
- (2) AD9219 samples radar analog video, and FPGA sends a 5M differential clock to AD9219. According to the FCO and DCO input by AD9219, 10 bit radar digital video sampled by AD9219 is obtained (see Fig. 4).

- (3) The purpose of the pixel solution module is to transform the polar coordinate system into a rectangular coordinate system. The process is: convert the azimuth code into the format required by CORDIC IP input, and obtain the sin and cos code values of the corresponding angle of the azimuth code. Taking the 5M bandwidth of 20 km and the range accuracy of 30m as an example, the effective sampling point is $20 \text{ km}/30 \text{ m} = 667$. Take the distance sweep pulse as the starting point, count 0 to 666, and use the divider to convert it to the pixel length of 0–511. Use shift addition to sin,cos and pixel length to obtain the projection of the pixel length in the X and Y directions at the current angle, and then calculate the address of the pixel. In this way, the radar video data `rgb_data`, data valid signal `rgb_Valid`, radar video data address `rgb_addr`.

$$\begin{aligned} x &= l * \cos \alpha + 512 \\ y &= -l \sin \alpha + 512 \quad (\alpha \in (-\pi, \pi)) \end{aligned} \quad (1)$$

3.2 Design of Afterglow

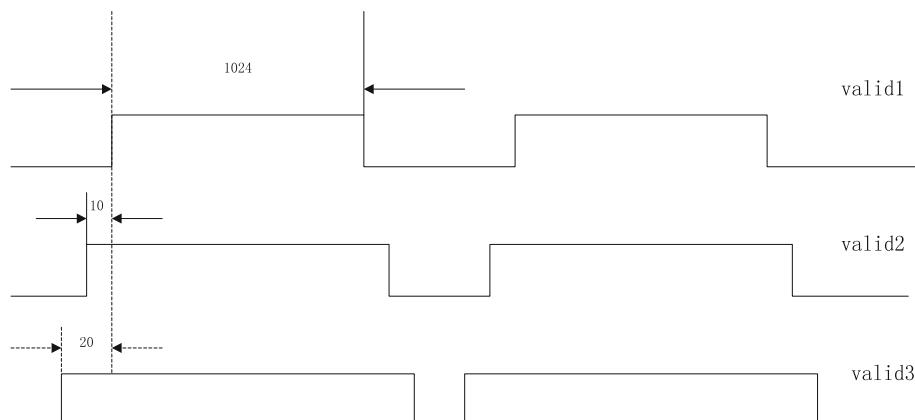


Fig. 5. Valid timing design

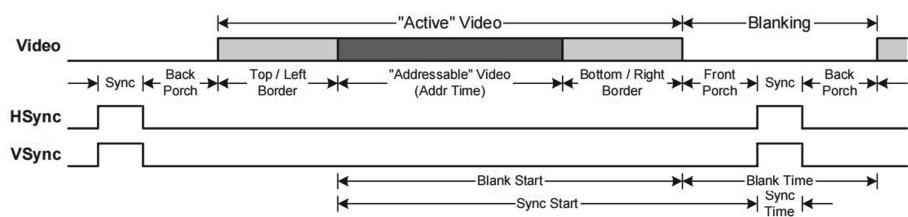


Fig. 6. Vesa timing design

Timing Name	=	1280 x 1024 @ 75Hz;			
Hor Pixels	=	1280;	// Pixels		
Ver Pixels	=	1024;	// Lines		
Hor Frequency	=	79.976;	// KHz	=	12.5 usec / line
Ver Frequency	=	75.025;	// Hz	=	13.3 msec / frame
Pixel Clock	=	135.000;	// MHz	=	7.4 nsec $\pm 0.5\%$
Character Width	=	8;	// Pixels	=	59.3 nsec
Scan Type	=	NONINTERLACED;		// H Phase =	6.9 %
Hor Sync Polarity	=	POSITIVE;	// HBlank	=	24.2% of HTotal
Ver Sync Polarity	=	POSITIVE;	// VBlank	=	3.9% of VTotal
Hor Total Time	=	12.504;	// (usec)	=	211 chars = 1688 Pixels
Hor Addr Time	=	9.481;	// (usec)	=	160 chars = 1280 Pixels
Hor Blank Start	=	9.481;	// (usec)	=	160 chars = 1280 Pixels
Hor Blank Time	=	3.022;	// (usec)	=	51 chars = 408 Pixels
Hor Sync Start	=	9.600;	// (usec)	=	162 chars = 1296 Pixels
// H Right Border	=	0.000;	// (usec)	=	0 chars = 0 Pixels
// H Front Porch	=	0.119;	// (usec)	=	2 chars = 16 Pixels
Hor Sync Time	=	1.067;	// (usec)	=	18 chars = 144 Pixels
// H Back Porch	=	1.837;	// (usec)	=	31 chars = 248 Pixels
// H Left Border	=	0.000;	// (usec)	=	0 chars = 0 Pixels
Ver Total Time	=	13.329;	// (msec)	=	1066 lines HT - (1.06xHA)
Ver Addr Time	=	12.804;	// (msec)	=	1024 lines = 2.45
Ver Blank Start	=	12.804;	// (msec)	=	1024 lines
Ver Blank Time	=	0.525;	// (msec)	=	42 lines
Ver Sync Start	=	12.816;	// (msec)	=	1025 lines
// V Bottom Border	=	0.000;	// (msec)	=	0 lines
// V Front Porch	=	0.013;	// (msec)	=	1 lines
Ver Sync Time	=	0.038;	// (msec)	=	3 lines
// V Back Porch	=	0.475;	// (msec)	=	38 lines
// V Top Border	=	0.000;	// (msec)	=	0 lines

Fig. 7. Vesa timing parameter

- (1) Count based on 135M pixel clock. As shown in Fig. 7, follow the standard VESA timing parameters such as the Back Porch, Front Porch, Sync etc. Then generate field synchronization signal Vsync, line synchronization signal Hsync and effective video signal de. the relationship between signals is shown in Fig. 6.
- (2) Use the IP Core of FPGA block ram to generate a dual port Bram with address depth of $1024 * 1024$ and bit width of 12 bit, which is used to store radar data of $1024 * 1024$ pixels. The reason why the depth is set to $1024 * 1024$ is that in the DVI video display with a resolution of $1280 * 1024$, the area of 1024 rows and 1024 columns on the left is the radar video display area. The bit width is 12 bit, which is limited by the resources of FPGA Bram, and RGB is 4 bit each.
- (3) Generate a FIFO with a bit width of 32 bit and a depth of 1024 for caching radar data with a bandwidth of 5M. 32 bit low 20 bit storage address and 12 bit high storage RGB data. Because the speed of reading FIFO (135M) is much higher than that of

writing FIFO (5M), so FIFO would not overflow. Here FIFO keeps the margin and the depth is set to 1024.

- (4) Generate field synchronization signals, valid1, and valid2 and valid3 signals according to VESA standard. Valid1 is the area of $1024 * 1024$, that is, a section of high-level clock with 1024 pixels, a total of 1024 sections. Valid2 is $1024 * 1044$ and valid3 is $1024 * 1064$. The sequence is shown in the figure. Valid2 lags behind valid3 by 10 pixel clocks, and valid1 lags behind valid3 by 20 pixel clocks (see Fig. 5).
- (5) FIFO read enable rd_en = ! valid3. When valid2 == 0, update the radar data of Bram. The reason for this design is that the operation of updating Bram's radar data and afterglow are staggered in time to prevent conflict.
- (6) Afterglow can be understood as a function of initial brightness value, attenuation rate and time. The time can be obtained from the field synchronization signal. The initial value of brightness can be read out by Bram reading port. The default attenuation rate is brightness minus 1 each time, which can be controlled by UART port message. When the frame count reaches a certain value, traverse the storage space of Bram, read out the data, subtract 1 from the data, and then write to Bram to complete the update of afterglow data.

3.3 Design of Local Amplification Function

- (1) After receiving the radar video sent by SRIO, FPGA A stores the radar video in the $1024 * 1024$ area on the left side of the image with a DDR3 resolution of $1280 * 1024$, and the address of each line is $0-1023, 1280-2303, 1280 * n - 1280 * n + 1023$ ($n = 0, 1, 2, \dots, 1023$). For the input DVI image, the $256 * 1024$ area on the right side is intercepted according to the field synchronization signal vsync and the video effective signal de, and stored in the $256 * 1024$ area on the right side of the image in DDR3, with the corresponding address of $1024-1279, 2304-2559, (1280 * n + 1024) - (1280 * n + 1279)$. Then read out in the timing of DVI.
- (2) After receiving the UART local amplification command, intercept the $128 * 128$ image area according to the address of the upper left corner given by the command, enlarge it twice according to the bilinear interpolation algorithm, interpolate first, and then interpolate between lines, and store the obtained results in the dual port RAM with depth of $256 * 256$ and bit width of 24 bit. According to the line field synchronization signal, the effective signal valid_valid_256 of the area with the size of $256 * 256$ in the upper right corner is generated. When valid_256 is high level, read dual port RAM to get the amplified local area. When valid_256 is low, use the original DVI data. This completes the splicing of locally amplified DVI data. In order to improve the image quality, the data is input into DDR3, and then read out and displayed in the timing of DVI.
- (3) On FPGA, data changes with the clock like pipelining, there will only be one data at the same time. If you need to get the data of an area at the same time, you need to use the cache. The data in the same row can be delayed by one clock with one register, and the data between different rows can only be cached through FIFO. Usually, a FIFO caches only one row of data. For example, when opening a $3 * 3$ window, two FIFOs are required. In FPGA, FIFO consumes block ram resources,

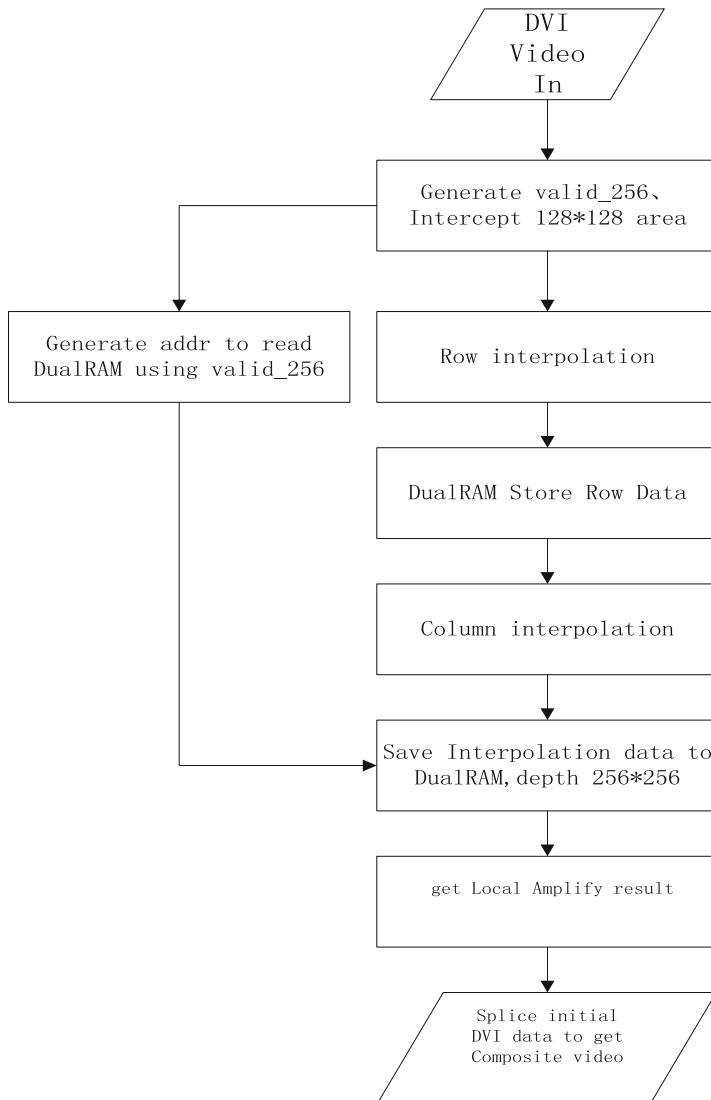


Fig. 8. Local amplification module

while the block ram of an FPGA is limited. Too much FIFO will inevitably affect the subsequent layout and wiring. Here, you can use dual port RAM instead of FIFO. A similar effect can be achieved by using a dual port RAM, which can save block ram resources.

There are two modes of RAM: first write and first read. The working mode of write first is to write data into RAM first and then read out the data in RAM. This mode obviously does not meet the requirements. When using dual port RAM, the

working mode of ram needs to be set to read first, which means that when writing new data to ram, the original data in RAM will be read out first.

If ram has three addresses, the original data a, b, b, e, f and g in RAM are the data to be written into RAM, as shown in Fig. 9:

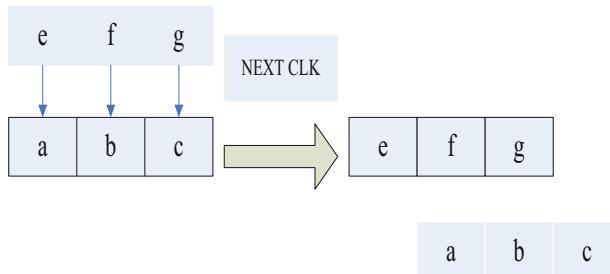


Fig. 9. Schematic diagram of RAM read first working mode

The number in the box in Fig. 8 represents the storage status of ram, and the number below the box represents the data in the original RAM (which can be regarded as the data pressed out of RAM). Since a, b and c need a clock to read from ram, when a, b and c are taken out, their addresses change from 0, 1 and 2 to 1, 2 and 0 (assuming that the address changes from 0 to 2, move in circles).

If e, f and g are a row of data corresponding to a, b and c, when e, f and g write addresses 0, 1 and 2, each column has a clock dislocation. Therefore, in order to align e, f and g with a, b and c, the addresses that should be written are 1, 2 and 0. For each row of data, the write address is increased by 1. In this way, the data of the corresponding window can be taken out at the same time.

If the image size is $128 * 128 * 8$ bit, the required dual port RAM size is 256 in depth and 8bit in width. Just connect the output of port A to port B, the clock (CLK) and the enable end (we) of port A are the same with port B. The address of port B is the address of port A increased by 128. Taking three lines of data as an example, Fig. 9 shows the changes and output of data in RAM. The shaded part indicates the final $3 * 3$ window. Port A and port B represent the data accessed in RAM, DATA_In indicates the data input to Port A of RAM.

In this way, the cached row data can be used to obtain the result of bilinear interpolation (Fig. 10).

- (4) Use MIG interface of IP Core to develop DDR3. The MIG interface is composed of a group of command ports, read ports and write ports, actually, more than one port is used, so the port of DDR3 needs to be expanded. Here, the port of DDR3 is expanded into two groups of ports, each of which contains two input ports and one output port. Each port has a FIFO to cache data, the FIFO of the write port is used to cache the input data, and the FIFO of the read port is used to cache the data read out from DDR3. The count value of the data in FIFO is introduced into

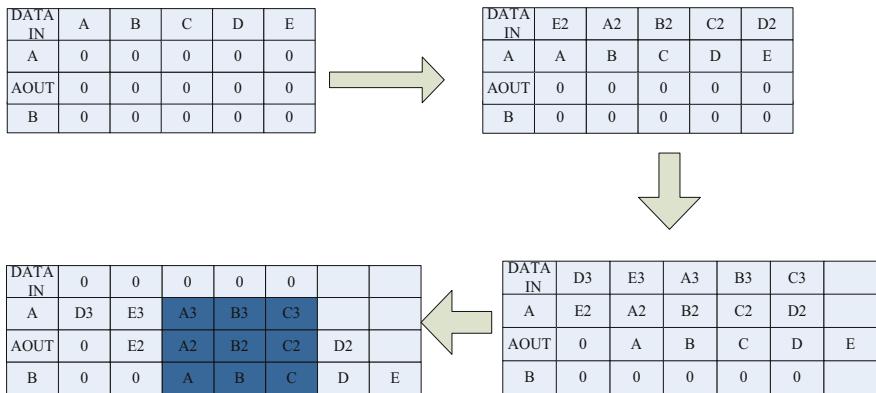


Fig. 10. Schematic diagram of using dual port RAM to cache row data

the arbitration port. The arbitration port determines which data occupies the user interface of DDR3 MIG according to the amount of data in FIFO (see Fig. 11).

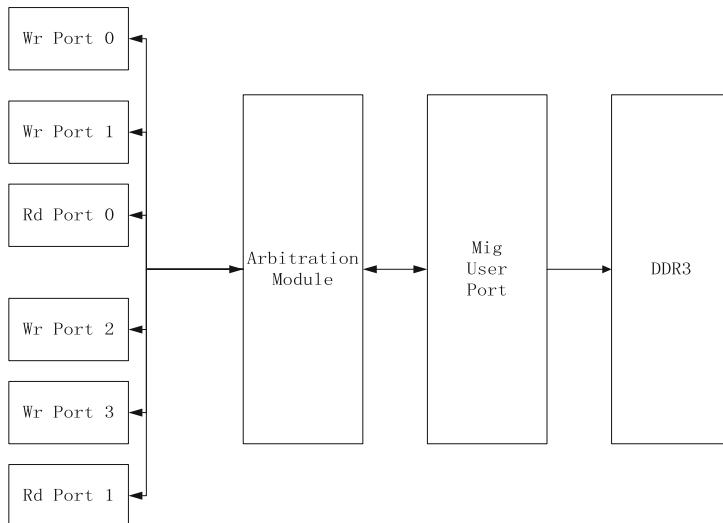


Fig. 11. Expansion diagram of DDR3 port

4 Experiment and Conclusion

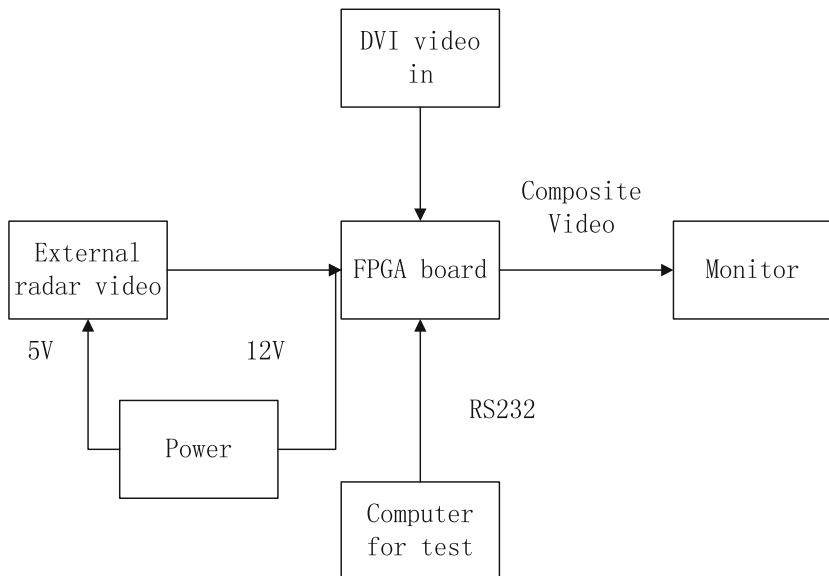
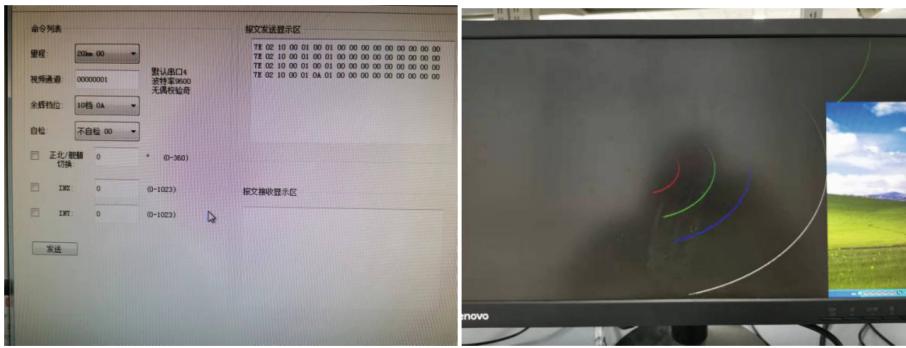


Fig. 12. Equipment for test



a) test interface

b) result

Fig. 13. Test interface and result

Radar analog video bandwidth 5M, DVI video resolution ratio 1280 * 1024 * 60 Hz, Uart Baudrate 9600, SRIO 3.125 Gbps 4X. The development tool of test software is Qt5.5, and the development tool of FPGA is vivado2017.4 (Figs. 12 and 13).

The hardware implementation method of radar video scanning transformation based on dual FPGA has the functions of bow/true north switching, range switching, self-inspection, afterglow and local amplification. The overall structure is simple, the decoupled architecture has high reliability and strong expansibility. For the realization of

afterglow, the on-chip Bram resources are used, which has strong hardware scalability, can reach 75 afterglow gears in theory, and the software design is simple.

References

1. Xiao, X., Lv, L.: Hardware design and implementation of a radar display technology. *Ship Electron. Eng.* **28**(7), 113–115 (2008)
2. Wang, X.M., Zhang, G.: Radar and Detection, p. 7. National Defense Industry Press, Beijing (2008)
3. Sun, B.: Design and implementation of radar signal processing algorithm based on FPGA. Beijing University of Technology, Beijing (2014)
4. Cao, Y., Yao, Y., et al.: Design of radar velocity measurement system based on TMS320F28335. *Electron. Device* **37**(1), 45–49 (2014)
5. Zhai, G., Ji, Y.: High performance raster radar display system based on DVI Technology. *Radar Countermeas.* **2**, 49–56 (2009)
6. Liu, C., Wen, D.: Long afterglow simulation of PPI radar on raster scanning display. *Comput. Simul.* **3**, 42–47 (2012)
7. Li, H., Zhu, X., Gu, C.: Development, design and application of Verilog HDL and FPGA, pp. 125–127. National Defense Industry Press, Beijing (2013)
8. Gao, Y.: Digital Signal Processing Based on FPGA. Electronic Industry Press, Beijing (2012)

Image Processing



An Image Binarization Segmentation Method Combining Global and Local Threshold for Uneven Illumination Image

Jin-Wu Wang^(✉), Daiwei Xie^(✉), and Zhenmin Dai^(✉)

The 723th Research Institute of China Shipbuilding Industry Corporation, Yangzhou 225001,
China

csicmarco@163.com, 892084639@qq.com, 421115959@qq.com

Abstract. For images with uneven illumination, the segmentation method based on global threshold will be affected by illumination, and the effect to low contrast and uneven illumination images is poor. In the segmentation method based on local threshold, the image will have discontinuous gray distribution at the boundary of different sub images, resulting in artifacts. In this paper, a binary image segmentation method is proposed, which uses the minimum filter to eliminate the uneven illumination and combines the global threshold with the local threshold according to the edge information of Canny operator. Experiments show that this method can achieve ideal segmentation effect for images with uneven illumination.

Keywords: Uneven illumination · Canny operator · Combination of global and local thresholds · Binarization

1 Background

Image binarization is the most widely used image segmentation technology. It is widely used in image processing such as automatic target recognition and target tracking. Most of the existing binarization methods belong to threshold methods, mainly including global threshold method and local threshold method.

Generally speaking, the global thresholding method is simple to implement and has obvious effect on the image with obvious bimodal histogram, but it has poor effect on the image with low contrast and uneven illumination, so the application range is greatly limited. The local threshold method can adapt to more complex situations and is more widely used than the global threshold method. However, it often ignores the boundary feature information of the image, which makes some different regions in the original image become a large region after binarization, resulting in the loss of some important information of the binarization result image.

For low contrast and uneven illumination images, using global threshold segmentation and local threshold segmentation is not enough to segment the target. In this paper, an image binarization algorithm is proposed, which eliminates the uneven illumination, combines Canny operator and combines global and local threshold.

2 Basic Threshold Segmentation Algorithm

Gray discontinuity occurs on the boundary between different regions in the image, the step change of gray forms the boundary of the region, so we can find the algorithm of color or gray mutation of adjacent pixels and segment according to the gray discontinuity of each pixel. Threshold segmentation is a method to segment the image into background and object according to the difference of image gray value. The core of threshold segmentation is the problem of threshold selection. According to different threshold selection methods, the main algorithms of image segmentation include histogram threshold method, iterative method, Otsu method and local threshold method.

2.1 Histogram Threshold Method

Histogram threshold method's threshold is mainly determined by analyzing the gray histogram of the image. An image is only composed of object and background, and its gray level histogram presents obvious bimodal values, as shown in the Fig. 1. For this kind of image, the gray value at the bottom of the valley between the bimodals is selected as the threshold to segment the image. However, an image with obvious bimodal characteristics can be said to be an ideal situation. In fact, it is difficult to find such an image. An image usually consists of multiple objects and background. At this time, its gray histogram may show multiple obvious peaks, so the gray value at the peak and valley between the peaks can still be taken as the threshold T . at this time, there are multiple thresholds to segment the image, which can become the choice of multi peak threshold. Assuming that there are three peaks, as shown in Fig. 2, the gray values T_1 and T_2 at the two peaks and valleys can be selected as the threshold.

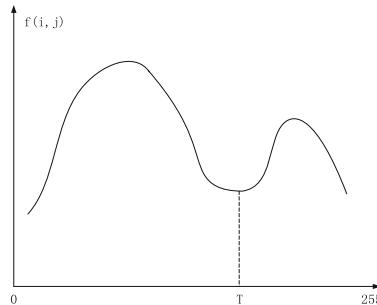


Fig. 1. Multimodal histogram distribution

$$f(i, j) = \begin{cases} 255, & f(i, j) \geq T \\ 0, & f(i, j) < T \end{cases} \quad (1)$$

$$f(i, j) = \begin{cases} 255, & T_1 \leq f(i, j) \leq T_2 \\ 0, & others \end{cases} \quad (2)$$

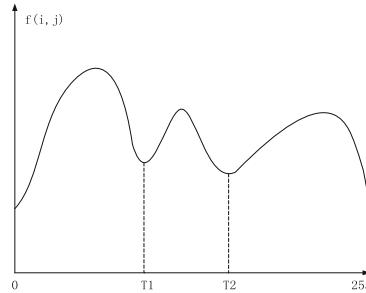


Fig. 2. Multimodal histogram distribution

2.2 Iterative Threshold Segmentation

Iterative method is also a method to select the appropriate threshold in the process of image segmentation. It is based on the idea of approximation, and uses the program to automatically calculate the more appropriate segmentation threshold through threshold iteration. Iterative method assumes a threshold in the initial condition, and updates the assumed threshold to obtain the best threshold through iterative operation of the image. Iterative threshold segmentation main algorithms:

- (1) The minimum gray value R_{\min} and the maximum gray value R_{\max} of the image are calculated, and the initial threshold is $T_0 = \frac{R_{\min} + R_{\max}}{2}$.
- (2) Divide the image into target and background according to the threshold, and calculate the average gray value of the two parts:

$$R_0 = \frac{\sum_{R(i,j) < T_K} R(i,j) \times N(i,j)}{\sum_{R(i,j) < T_K} N(i,j)} \quad (3)$$

$$R_G = \frac{\sum_{R(i,j) > T_K} R(i,j) \times N(i,j)}{\sum_{R(i,j) > T_K} N(i,j)} \quad (4)$$

$R(i, j)$ is the gray value of (i, j) points on the image, $N(i, j)$ is the number of $R(i, j)$. T_K is the threshold.

- (3) Reselect the threshold $T_{K+1}, T_{K+1} = \frac{R_0 + R_G}{2}$
- (4) Cycle steps 2–3 until $T_K = T_{K+1}$, the image can be segmented with the best threshold.

2.3 Otsu Segmentation

Otsu method is an adaptive threshold determination method, also known as maximum interclass variance method, or Ostu for short. This method is based on the separability of target and background in the image. It is based on the assumption that the mixed density function composed of target and background in the image is composed of two

sub distributions subject to equal variance normal distribution. For image $I(x, y)$, the segmentation threshold foreground (i.e. target) and background is T , the proportion of foreground points in the whole image is ω_0 , its average gray level is μ_0 ; the proportion of background points in the whole image is ω_1 , its average gray level is μ_1 ; The total average gray level of the image is ω , The variance between classes is recorded is g .

$$\mu = \omega_0 \times \mu_0 + \omega_1 \times \mu_1 \quad (5)$$

$$g = \omega_0 \times (\mu_0 - \mu)^2 + \omega_1 \times (\mu_1 - \mu)^2 \quad (6)$$

Assuming that the background of the image is dark and the size of the image is $M \times N$, the number of pixels in the image whose gray value is less than the threshold T is recorded as N_0 , and the number of pixels whose gray value is greater than the threshold T is recorded as N_1 , then:

$$\omega_0 = \frac{N_0}{M \times N} \quad (7)$$

$$\omega_1 = \frac{N_1}{M \times N} \quad (8)$$

$$N_0 + N_1 = M \times N \quad (9)$$

$$\omega_0 + \omega_1 = 1 \quad (10)$$

$$g = \omega_0 \times \omega_1 (\mu_0 - \mu_1)^2 \quad (11)$$

When the variance between classes is maximal, the threshold T can be obtained by ergodic method.

2.4 Local Threshold Method

There are shadows, uneven illuminance, different contrast and background gray changes in the image, at this time, if only a fixed global threshold is used to segment the whole image, the segmentation effect will be affected because it can not take into account the situation of all parts of the image. So, the local threshold method is adopted, a set of thresholds related to the pixel position are used to separate each part of the image division. The simplest method is to divide the image into several small images, segment each sub image by threshold method, and then merge the segmented small areas together to obtain the complete segmentation result of the whole image.

Bernsen method, Chow and Kanekos Method, Eikvil Method, Mardia and Hainsworth's Method are Common methods.

3 Remove Uneven Illumination

In the process of video and image acquisition, the influence of complex environment, mutual occlusion between objects and variable ambient lighting, would lead to uneven lighting of the scene. As a result, there is too much strong light in the bright area of the image.

The lack of illumination in the dark area leads to some important details that cannot be highlighted or even covered up, which seriously affects the visual effect and application value of the image. Therefore, it is a hot topic in the field of image processing to do research on the correction of uneven illumination image and eliminate the influence of uneven illumination on the image [1].

Due to the existence of uneven illumination, the brightness of the target is similar to background during binarization, so the target cannot be segmented from the background (Fig. 3).

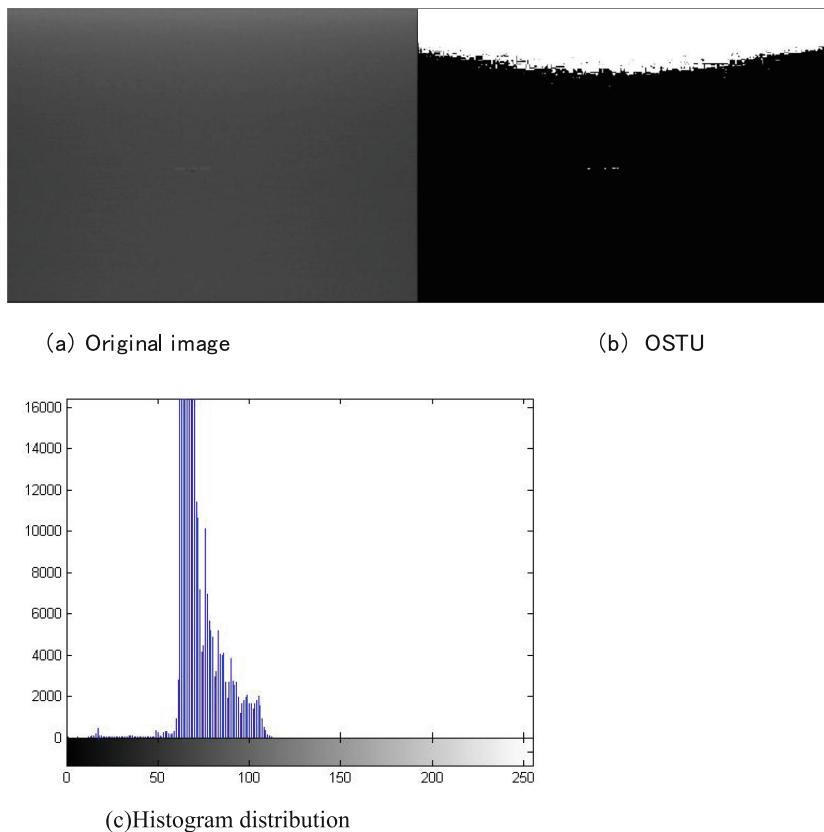


Fig. 3. Uneven illumination image and its segmentation effect by Ostu method

As shown in (a), at the top of the picture, the pixel brightness is too large due to the influence of illumination. After direct segmentation using Otsu method, the top of

the image is white, and the target in the middle of the image can not be completely segmented.

Aiming at the problem of uneven illumination, some scholars use morphological methods to deal with it, and use a disk-shaped structural element with a radius of 40 for opening operation. This structural element is large enough to not fit any object. After operation, only an approximate background is left. Subtracting the image from the original image (i.e. top hat transformation) will make the background more uniform [2]. However, due to the large size of structural elements, this method has a large amount of computation and slow operation speed. It is not applicable in occasions with high real-time performance. The method based on Retinex theory has color constancy, but this kind of method will produce halo phenomenon where the brightness of the image changes suddenly [3, 4]. The unsharp mask method decomposes the image into high-frequency components and low-frequency components for processing respectively, but it is difficult to find the optimal threshold accurately between high-frequency and low-frequency in practical application, and it's also difficult to take the balance between detail enhancement and naturalness into account [5, 6]. The method based on spatial variable illuminance map uses the illumination distribution characteristics of the scene to correct the image, but the illumination component solved by the method of single-scale Gaussian function has some problems, such as poor expressiveness of illumination detail information [7, 8].

The image processed in this paper is black-and-white image, which has high requirements for real-time performance. Therefore, this paper uses a high real-time algorithm.

3.1 Illumination Reflection Imaging Model

According to the imaging principle, the image formed in the visible light range is generated after the light emitted from the object surface in the scene reaches the imaging unit. Generally, a digital image can be regarded as a two-dimensional function $f(x, y)$, and the value of the function is the image brightness value at the coordinate point (x, y) . It is composed of the product of the illumination component $i(x, y)$ incident into the scene and the reflection component $r(x, y)$ on the object surface [9], and its expression is as follows:

$$f(x, y) = i(x, y) * r(x, y) \quad (12)$$

This model is called illuminance reflection imaging model, and its spatial relationship is (Fig. 4):

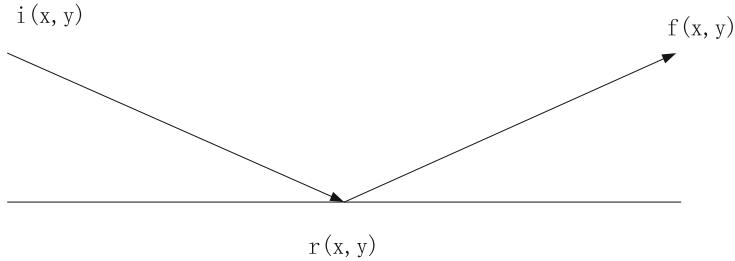
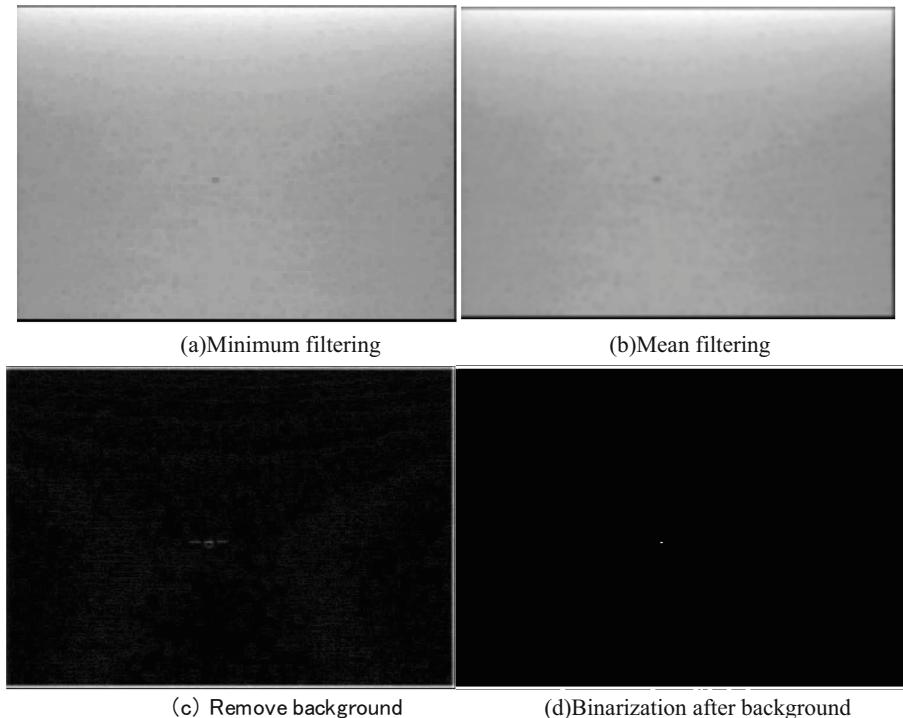


Fig. 4. Light reflection imaging model

In the illumination reflection imaging model, the illumination component represents the low-frequency characteristics of the image, while the reflection component reflects the high-frequency details of the image, which determines the essential characteristics of the image. For the image with uneven illumination, due to the uneven distribution of illumination components in the scene, the brightness value of the image in the area with strong illumination is too strong, while the brightness value of the image in the area with weak illumination is insufficient, which not only reduces the visual quality of the image, but also makes some important details unable to be extracted.

3.2 Extraction of Illumination Component Based on Minimum Filtering

According to Retinex theory, the following assumptions are made: the illumination component of the real scene image mainly exists in the low-frequency part of the image and changes gently as a whole; The reflection component mainly exists in the high-frequency part of the image, such as edge, texture, etc., and its change is relatively violent [9]. Therefore, it is hoped that the extracted lighting component of the scene only contains the lighting change information, not the detailed information of the image, so as to meet the assumption conditions of the lighting component of the scene better. Therefore, this paper uses minimum filtering to obtain the preliminary illumination map, then uses mean (or Gaussian) filtering to obtain the final illumination distribution map, and finally subtracts the illumination map from the original image to obtain the foreground target.



removal

Fig. 5. Effect diagram of eliminating uneven illumination

As can be seen from Fig. 5, after removing the background, the overall brightness of the image is uniform, and there is no wrong white areas after binarization at the top. According to the histogram distribution of the target, it can be seen that the histogram has no bimodal characteristics, and the segmentation effect is poor when using global threshold segmentation (Fig. 6).

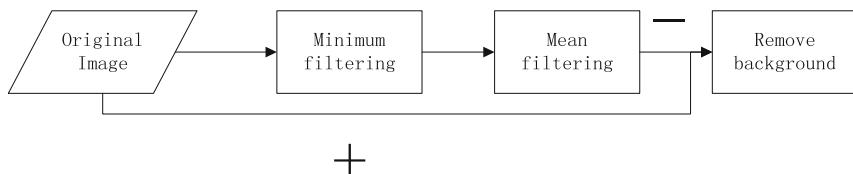


Fig. 6. Flow chart of eliminating uneven illumination

4 Global and Local Threshold Fusion Combined with Canny Operator

Compared with Sobel, Prewitt, Roberts, Laplacian and log operators, Canny operator has more advantages with the advantages of low error rate, better positioning of edge points and single edge response [2], and can better detect the target edge.

Canny can be implemented in four steps:

First step, image denoising. The gradient operator can be used to enhance the image, which is essentially realized by enhancing the edge contour. However, they are greatly affected by noise, So the first step is to remove the noise. The noise is where the gray changes greatly, so it is easy to be identified as a pseudo edge. Use gaussian filter to smooth image. Actually, denoising and edge detection are a pair of contradictions, so use the first derivative of Gaussian function to obtain the best balance between them.

$$f_s = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} * f(x, y) \quad (13)$$

The second step is to calculate the image gradient and obtain the possible edges. This step can only get the possible edge. Because the place where the gray level changes may or may not be the edge. In this step, there is a set of all possible edges.

$$\Delta f_s = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right]^T = [g_x, g_y]^T \quad (14)$$

For Images are two-dimensional discrete data,

$$\begin{aligned} g_x &= \frac{f(x+1, y) - f(x, y)}{(x+1) - x} = f(x+1, y) - f(x, y), \\ g_y &= f(x, y+1) - f(x, y) \end{aligned} \quad (15)$$

$$M(x, y) = \sqrt{g_x^2 + g_y^2} \quad (16)$$

$$\theta = \tan^{-1} \frac{g_y}{g_x} \quad (17)$$

$M(x, y)$ means amplitude of gradient, θ means angle of gradient.

The third step, suppression of non maximum suppression. Generally, the places where the gray level changes are relatively concentrated. Keep the place where the gray level change the most violent in the gradient direction. Cut down the others. In this way, most of the points can be eliminated. Turn an edge with multiple pixels wide into a single pixel wide edge. It seems that ‘fat edge’ becomes ‘thin edge’.

The fourth step is hysteresis thresholding. Use double threshold to select edges. After non maximum suppression, there are still many possible edge points. Then set a double threshold, a low threshold and a high threshold. If the gray change is greater than the higher, it is set as strong edge pixels. If it is lower than the lower, it is eliminated. The edge between low and high threshold is set to weak edge. If there are strong edge pixels in its field, keep it, if not, eliminate it. The reason why use double threshold is when

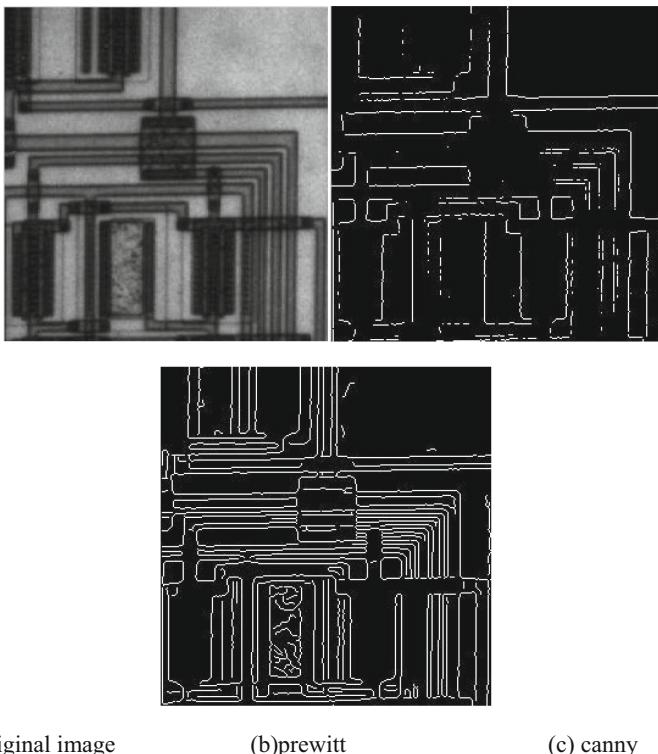


Fig. 7. Effect contrast of canny and prewitt

using single threshold it will generate false edges when the threshold is low and generate lose edges when the threshold is high.

As Fig. 7 shows, canny operator has the advantages of low error rate, better positioning of edge points and single edge response. The edge of canny is thinner and more accurate than edge of prewitt. So use canny operator to find the edges of target.

According to the edge [10] obtained by Canny operator, calculate the upper, lower, left and right coordinates of the target, intercept the local image in this coordinate, and binarize it with Otsu method. The processed segmented image can be obtained by logical or operation between the local image binarization result and the global image result (Figs. 8 and 9).

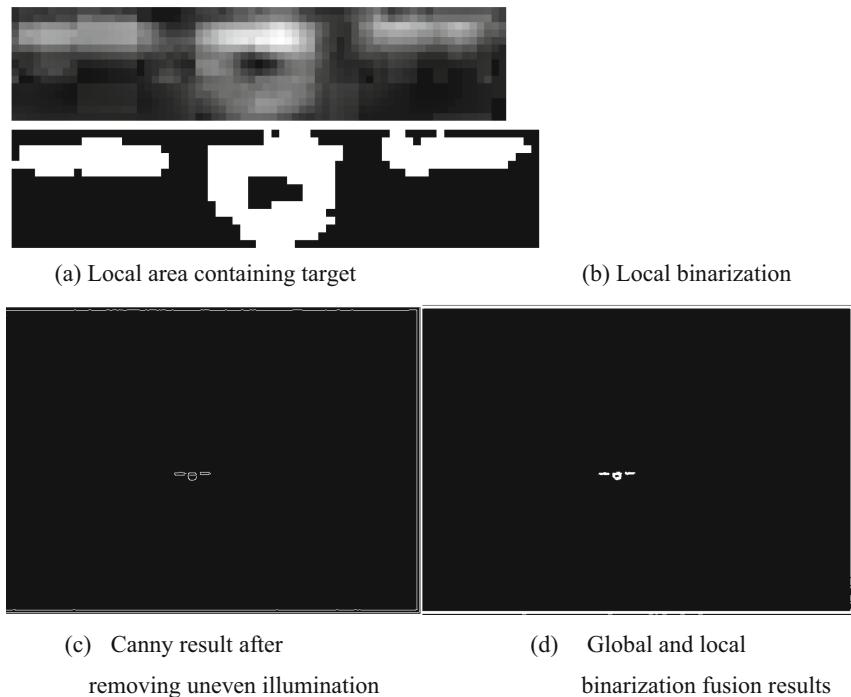


Fig. 8. Effect diagram of combining Canny operator, global and local binarization results

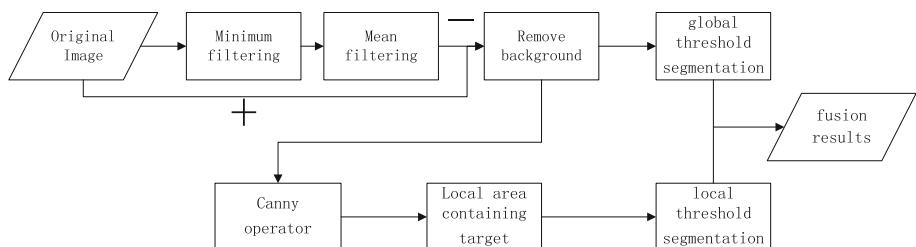


Fig. 9. Schematic diagram of global and local binarization algorithm combined with Canny operator to eliminate uneven illumination

5 Experiment and Conclusion

For the uneven illumination image, the global threshold segmentation can not segment the target. In this paper, an image binarization segmentation method based on the combination of global and local thresholds for uneven illumination image is proposed. This method uses the minimum filter to eliminate the phenomenon of uneven illumination. By detecting the target edge with Canny operator and the combining global and local binarization, the target of low contrast and uneven illumination image can be segmented, and the ideal effect is achieved.

References

1. Liu, Z.C., Wang, D., Liu, Y., et al.: Adaptive correction algorithm for uneven illumination image based on two-dimensional gamma function. *J. Beijing Univ. Technol.* **36**(2), 191–196 (2016)
2. Rafael, C., Richard, E.: Digital Image Processing, 3rd edn. Publishing House of Electronics Industry, Beijing (2011)
3. Md Shukri, D.S., Asmuni, H., Othman, R.M., et al.: An improved multiscale Retinex algorithm for motion-blurred iris images to minimize the intraindividual variations. *Pattern Recogn. Lett.* **34**, 1071–1077 (2013)
4. Li, J.: Application of image enhancement method for digital images based on Retinex theory. *Optik* **124**, 5986–5988 (2013)
5. Kareem, S., Kale, I., Morling, R.C.S.: Automated malaria parasite detection in thin blood films: a hybrid illumination and color constancy insensitive, morphological approach. In: Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems, pp. 240–243. IEEE (2012)
6. Maragos, P., Pessoa, L.F.C.: Morphological Filtering for Image Enhancement and Detection, the Image and Video Processing Handbook. Academic Press, Harcourt (1999)
7. Lee, S., Han, H., Kwak, B., et al.: Color image enhancement method using a space variant luminance map. In: Proceedings of the 2010 Digest of Technical Papers International Conference On Consumer Electronics [S.I.], pp. 413–414. IEEE (2010)
8. Lee, S., Han, H., Kwak, B., et al.: A space variant luminance map based color image enhancement. *IEEE Trans. Cons. Electron.* **56**(4), 2636–2643 (2010)
9. Land, E.H.: An alternative technique for the computation of the designator in the retinex theory of color vision. *Proc. Natl. Acad. Sci.* **83**(10), 3078–3080 (1986)
10. Wang, T., Xu, Y., Kang, H.L., et al.: Image binarization method combined with Canny operator. *Microcomput. Appl.* **26**(2), 4–7 (2010)



Optimization of Vessel Segmentation Using Genetic Algorithms

Jared Cervantes, Dalia Luna, Jair Cervantes^(✉) , and Farid García-Lamont

Fraccionamiento el Tejocote, Autonomous University of Mexico State, Av. Jardín Zumpango s/n, 56255 Texcoco, Mexico
jcervantesc@uaemex.mx

Abstract. Vessel segmentation is an important task to extract helpful information from retinal images that can help make a retinopathy diagnosis. A good segmentation perfectly represents the structure and obtains patterns that diagnose retinal diseases. Most of the current methods require many parameters, and the final quality of vessel segmentation depends on these parameters, which increases the complexity of the methods. We propose a new Vessel segmentation algorithm to address these issues using genetic algorithms. The method uses several steps to segment the retinal images. However, each of the parameters used in the steps is optimized by the genetic algorithm. To evaluate the performance of the proposed method, we achieved experiments with two freely accessible datasets for vessel segmentation, digital retinal images for vessel extraction (Drive) and the Child Heart Health Study in England (Chase-db1). Experimental results show an acceptable performance of the proposed method using sensitivity (0.7941), specificity (0.9451), and accuracy (0.9578) performance metrics.

Keywords: Vessel segmentation · Genetic algorithm · Optimal segmentation

1 Introduction

The blood vessel network is one of the most important tools for detecting retinal vein occlusion, diabetic retinopathy, and early diagnosis of glaucoma. The blood Vessel network provides valuable information for the analysis of diseases [1]. Vessel segmentation is an important task in retinal image analysis. There are several techniques in the literature for automatic segmentation of retinal images.

Most of the current research on Vessel segmentation uses several steps to find the correct segmentation. However, each step can use one or several methods and thus, several parameters must be optimized. The correct optimization of the parameters allows for obtaining a good segmentation. Identifying the right processes for optimal segmentation is not an easy process. It requires, in some cases, a deep knowledge of each process that is only obtained through years of experience. That is why the development of automatic segmentation techniques is essential. In addition to the above, obtaining the processes necessary for optimal segmentation is not easy since the combination of processes in

most cases is vast. In addition to the above, there is no general solution for all segmentation cases. That is why evolutionary algorithms are necessary and justified due to the enormous solution space. Evolutionary algorithms have become an essential tool for solving optimization problems. Currently, there are several algorithms to optimize parameters. However, these are designed for some specific segmentation techniques. In [2] the authors use a basic genetic algorithm to optimize parameters of the FCM clustering algorithm, while in [3] the authors use a particle swarm-based algorithm to find segmentation boundaries. In this article, we use three different segmentation algorithms with different preprocessing algorithms and morphological operators.

This paper presents a new algorithm to optimize vessel segmentation. The proposed algorithm uses genetic algorithms to optimize the parameters of each step in the segmentation. The rest of this paper is organized as follows. Section 2 shows a review of the research on vessel segmentation. Section 3 describes the proposed method. Section 4 presents the experimental results obtained using the proposed method and some comparisons. Section 5 concludes the paper with the conclusions and future research directions.

2 State of the Art

The early detection of signs of retinopathy is a problem that has been pursued in recent years, where the process of blood vessel segmentation is essential. This segmentation provides crucial information for diagnosis, planning, and treatment. Methods and techniques have stood out in solving this problem. Some, such as [4–6], are based on an encoder-decoder architecture. In [4], a deep neural network called a multi-scale dense network (MD-Net) is used to segment fundus retinal blood vessels by using multi-scale information by a residual atrous spatial pyramidal pooling (Res-ASPP) and dense multilevel fusion to minimize feature loss. While in [5], the method is composed of a contextual information-enhanced dilated convolutional network (CIEU-Net) containing 47 connected layers, employing precise segmentation and integrating the cascaded dilated modules with multiple grid strategy of the pyramidal module with spatial continuity.

In [6], new edge-aware flows are introduced to guide retinal vessel segmentation, making segmentation more sensitive to fine capillary edges. In [7, 8], the authors work with a segmentation method supported by the Contrast Limited Adaptive Histogram Equalization (CLAHE) process in the preprocessing stage, followed by a median filter to reduce noise in the preprocessed image. Then, a vector is extracted based on pixels of the preprocessed images, and finally, this vector is given as input to the Artificial Neural Network (ANN). Five different feature groups are used for feature extraction. These functions are edge detection, morphology, statistics, gradient, and Hessian matrix. We can also find that [8] uses edge detection filters to extract the feature vector; the found features are used to train an artificial neural network to recognize each pixel as belonging to blood vessels or not.

In [9], a method is handled that is composed of four steps; in the first step, the initial color images are converted to gray with predetermined weights to increase the contrast of the image. Second, the image intensities are extended from the regions of interest to the entire image domain with a doubling operation to avoid introducing unwanted

boundaries by existing image filtering operations in the next step. Third, an improved multi-scale method inspired by Frangi filtering is proposed to enrich the image contrast between blood vessels and other objects that may exist in the image. Finally, an enhanced level set model is proposed to segment blood vessels from the enhanced and original gray images. In [10], the authors use a convolution neural network. The CNN model can perform the mapping relationship from the gray fundus scale to the discriminant matrix. This model achieves the initial segmentation of retinal vessels. The prediction of uncertain pixels is checked using the geometric features of the vessels and through the analysis of connected zones.

Currently, there are different algorithms based on CLAHE for vessel segmentation. The different algorithms are used to obtain good precision with the metrics used. However, the user must manually optimize various parameters to achieve the best segmentation in most cases. This paper uses different steps to obtain vessel segmentation, and each step contains its parameters. Optimizing these parameters is not an easy task, so a specialized method is required. This paper presents a new method that automatically uses genetic algorithms to optimize the parameters and automatically obtain the best vessel segmentation.

3 Proposed Method

This section describes the steps used with the proposed method. The main objective of this research is to optimize vessel segmentation by using the CLAHE algorithm.

Table 1. Parameters optimized with the proposed algorithm.

Method	Parameter	Interval
CLAHE	Distribution	Uniform, Rayleigh, Exponential
	Alpha (Rayleigh an exponential)	[0.2, 0.5]
	Tiles	8, 16, 32, 64
	Clip limit	[0.02, 0.05]
	Bins Histogram	128, 256
Filter	Gaussian (size and σ)	3, 5, 7 and $\sigma \in [1, 4]$
Segmentation	Algorithm	Otsu, Adaptative, k-means
Morphological operators	Erosion, Dilation, Closing and Opening	Order

The proposed algorithm considers the optimization of the parameters of the different steps of the algorithm. Table 1 shows the parameters used for each of the stages of the proposed algorithm. In this paper, three segmentation methods are considered to investigate the impact of their efficiency.

3.1 Datasets

In order to test the performance of the proposed method, we use ChaseDB1 and Drive datasets that are publicly available for blood vessel extraction research. The Drive

database contains 40 color eye fundus images available on the “Grand Challenge” website (<https://drive.grand-challenge.org/>). The resolution of the images is (565×584) . The database is divided into a test set (20 images) and a training set (20 images). The Chasedb1 dataset contains 28 pairs of images with a resolution of (999×960) . It is available on Kaggle (<https://www.kaggle.com/datasets/khoongweihao/chasedb1>). The dataset contains 21 pairs for training and seven pairs for validation. Both databases provide two ground truths manually marked by two independent observers. All the results presented in this paper are obtained using the first human observer in the database as ground truth. In our experiment, we used the training and testing image given in the dataset.

3.2 G Channel

Most of the methods in state of the art using CLAHE work on a green channel for contrast enhancement of the fundus image. Using only one channel reduces the size of the image by excluding the other two channels. However, using a single channel can lead to the loss of information. However, some researchers argue that the highest contrast between the vessels and the background can be observed in the green channel of the fundus images [11]. In our experiments, we use only the G channel.

3.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Contrast Limited Adaptive Histogram Equalization is a practical algorithm for low contrast enhancement of an image proposed in [12]. CLAHE provides greater precision in the contrast details of the image by performing noise removal. In recent years CLAHE has been used by many researchers because its use significantly improves images by generating sharper edges. CLAHE makes full use of the available grey-level spectrum, improving the local contrast.

3.4 Gaussian Filter

The pre-processing of the images helps to obtain better results during the segmentation. Using filters is an essential step in Vessel segmentation. The pre-processing can eliminate noise in the image by applying filters, like the mean filter or Gaussian filter. In our experiments, we consider Gaussian filters before implementing the segmentation techniques. The parameters to optimize the Gaussian filter in this research were σ and the size of the filter S_f .

3.5 Segmentation Techniques

In the digital image process, the step to segment an image refers to obtaining the region or shape of interest from the entire image. There are several methods to achieve this task; the most common technique is to find the threshold value where the intensities from the region of interest differ from the rest of the image [13]. The adaptive threshold is one of the primary automatic methods most used. To begin is computing the mean grey level of the

whole image, and then the threshold is refined iteratively until the threshold value stops changing. The algorithm is straightforward, a Low-Cost Computational Method and easy to implement. Otsu's method [14] selects an adequate threshold of a grey level to identify the areas of interest from the rest of the image or background. The search for a specific threshold minimizes the intraclass variance and maximizes the variance concerning the opposite frontier. On the other hand, K-means is an unsupervised algorithm that finds certain groups based on some similarity in the data with the number of groups K .

3.6 Morphological Operators

Morphological operations are mainly used to reconstruct the shapes of objects even though they have distorted shapes. These operations are based on simplifying the images using subsets of coordinates. There are different types of morphology: dilatation, erosion, opening, and closing. The primary morphological operations are erosion and dilation. From them, we can compose the opening and closing operations.

3.7 Parameters Optimization by GA

The steps of the genetic algorithm to optimize parameters of vessel segmentation are as follows:

1. Initial population. Initialize chromosomes, population with n individuals is defined by $S^g = \{s_1^g, s_2^g, \dots, s_n^g\}$, where g represents the current generation. The purpose of initializing the population randomly is to generate multiple feasible paths. The initial population of 80 individuals was created with ten variables. It is, $S^g = \{s_1^g, s_2^g, \dots, s_{80}^g\}$ and $s_i^g = [s_{i,1}^g, s_{i,2}^g, \dots, s_{i,10}^g]$, Where j in $s_{i,j}^g$ represents the parameters of Distribution, Alpha, Tiles, Clip limit, Bins, Size of filter, Sigma, Segmentation algorithm and type of morphological operation.
2. Fitness. Each chromosome in $s_i^g = [s_{i,1}^g, s_{i,2}^g, \dots, s_{i,10}^g]$ contains the necessary variables to get vessel segmentation. The variables define the parameters of CLAHE preprocessing, the segmentation technique and morphological operation. Each combination of these variables obtains a segmented image with a particular segmentation quality. The quality of this segmentation is compared with the optimal segmentation (*Ground truth*). We use the Accuracy metric to compare the segmentation obtained and the optimal segmentation. An individual will be more likely to survive the greater the fitness (accuracy).
3. Selection. Individuals are selected from the population using the roulette selection method. In this method, the selection probability of each individual is proportional to its fitness value, and the higher the fitness value, the higher the probability of being selected.
4. Crossover. This mechanism is essential in genetic algorithms and defines how genetic information passes from parents to children. The crossover fuses two individuals' genetic information and passes it on to the children. There are different crossover methods in the literature. In our experiments, we use two-point crossover.

5. Mutation. In genetic algorithms, mutation probability is much more significant than in any other operator. Mutation allows a chromosome to randomly modify its chromosome by making random jumps into the unknown. This operator is fundamental to the genetic algorithm, as it avoids getting trapped in local minima. In our experiments, we use mutation probability = 0.1.
6. The proposed algorithm finishes after a fixed number of iterations. In the last iteration, the string with the highest precision is stored with the obtained precision.
7. Elitism is generally used in GA to avoid losing or eliminating the individual with the best fitness. This technique selects the fittest individual or fittest individuals and passes them on intact to the next generation. The individual with the best aptitude is obtained and compared with the previous generation in each generation. If a chromosome is not obtained that improves the fitness obtained so far, the previous chain remains intact in the new generation.

4 Experimental Results

In order to test the performance of the proposed method we use some well-established vessel segmentation datasets and compare the proposed method with the state-of-the-art methods. This section shows the process used during the experiments and the metrics used to validate the results obtained.

4.1 Performance Metrics

Sensitivity, specificity, and accuracy are the quantitative measures used for evaluating this method; the measures are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where True Positive (TP) defines the total number of positive pixels which are correctly predicted; False Positive (FP) defines the total number of negative pixels which are incorrectly predicted, and True Negative (TN) is defined by the total number of negative pixels which are correctly predicted; False Negative (FN): the total number of positive pixels which are incorrectly predicted.

The performance measure for segmentation of retinal vessel structure is given in Tables 2 and 3. Table 2 shows the results obtained with Chasedb1 dataset and 3 shows the results obtained with Drive dataset. The results in Tables shows that the proposed method is competitive in comparison with the methods in the state of the art.

In Tables 2 and 3, it can be seen that the proposed algorithm fails in the results obtained in the Chasedb1 dataset. Compared to the other methods, the algorithm proposed in [5] has better results in both Accuracy and Sensitivity, although [5] does not report the results in specificity. On the other hand, the proposed algorithm improves on the others in Sensitivity in the Drive dataset. However, several algorithms reported better results in Accuracy and Specificity.

As can be seen in Fig. 1, the proposed algorithm has difficulties maintaining the thinnest lines. It could be due to the combination of morphological operators in the last phase and because morphological operators have some fixed parameters such as operator size. Future work could also optimize the operators' size, which could improve the results presented in this paper.

Table 2. Experimental results obtained and comparisons with other algorithms in the state of the art (Chasedb1).

Method	Accuracy	Sensitivity	Specificity
Azzopardi et al. [15]	0.938	0.758	0.958
Fraz et al. [16]	0.946	0.722	0.971
Proposed Method	0.9627	0.7847	0.9524

Table 3. Experimental results obtained and comparisons with other algorithms in the state of art (Drive).

Method	Accuracy	Sensitivity	Specificity
Vega et al. [17]	0.9412	0.7444	0.9600
Liskowski and Krawiec [18]	0.9472	0.7819	0.9748
Zhang et al. [6]	0.9476	0.7743	0.9725
Yang et al. [19]	0.9542	0.7653	0.9818
Li et al. [20]	0.9527	0.7569	0.9816
Proposed Method	0.9578	0.7941	0.9451

5 Conclusion and Discussions

Optimizing parameters of the techniques used in the process improves the quality of vessel segmentation. Good vessel segmentation improves the quality of the structure obtained. It results in the quality of the scrutiny of the shapes of the retina. In this paper, we optimize the parameters of the segmentation process. It improves vessel segmentation quality, which is reflected in the metrics used.

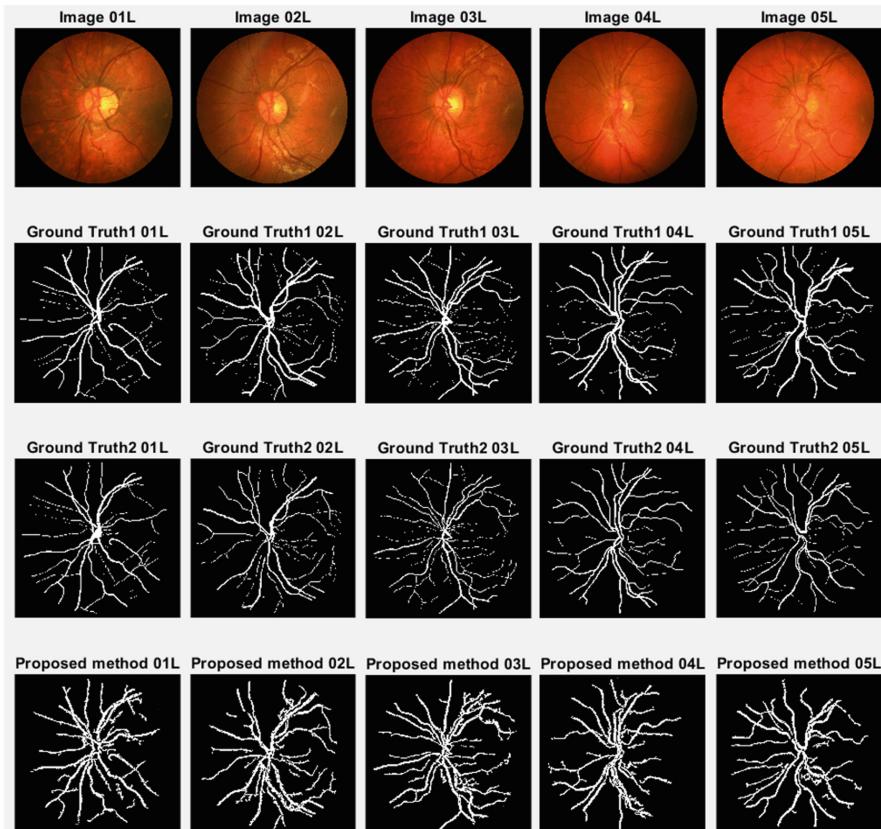


Fig. 1. Experimental results of vessel segmentation proposed. First row: Original fundus images; Second row: Ground truth 1; Third row: Ground truth 2; Fourth row: Proposed method

The use of equalization through the CLAHE algorithm enhances the contrast, which helps to improve the quality of segmentation. However, the fine adjustment of the CLAHE parameters is a very complex task since it requires the adjustment of several parameters, to mention this algorithm. In most algorithms like the one proposed, the complete segmentation process contains several steps with parameters that must be adjusted. This article presents a vessel segmentation algorithm based on genetic algorithms. Genetic algorithms allow optimization of each parameter used in the segmentation process from the CLAHE algorithm, the filter to be used and its parameters,

the segmentation method, and the morphological cleaning operators to be used. All the steps together improve the quality of segmentation.

The results obtained show that the proposed algorithm is competitive with the algorithms in state-of-the-art. Due to the use of the genetic algorithm, the proposed method can find the combination of parameters capable of detecting thin vessels and provides a precision of 0.9578 for the Drive dataset and 0.9578 for Chasedb1. The most significant disadvantage of the proposed algorithm is the difficulty in finding very thin vessels. Future work could also optimize the operator's size and combination of morphological operators, which could improve the results presented in this article.

References

1. Saffarzadeh, V.M., Shadgar, B., Osarch, A.: Vessel segmentation in retinal images using multi-scale line operator and K-means clustering. *J. Med. Signals Sens.* **4**(2), 122–129 (2013)
2. Xie, S., Nie, H.: Retinal vascular image segmentation using genetic algorithm plus FCM clustering. In: 2013 Third International Conference on Intelligent System Design and Engineering Applications (2013)
3. Ella Hassanien, A., El-bendary, N., Fahmy, A., Hassan, G.: Blood vessel segmentation approach for extracting the vasculature on retinal fundus images using Particle Swarm Optimization. In: 2015 11th International Computer Engineering Conference (ICENCO) (2015)
4. Shi, Z., Wang, T., Huang, Z., Xie, F., Liu, Z., Wang, B., Xu, J.: MD-Net: a multi-scale dense network for retinal vessel segmentation. *Biomed. Sig. Process. Control* **70**, 102977 (2021)
5. Sun, M., Li, K., Qi, X., Dang, H., Zhang, G.: Contextual information enhanced convolutional neural networks for retinal vessel segmentation in color fundus images. *J. Vis. Commun. Image Represent.* **77**, 103134 (2021)
6. Zhang, Y., Fang, J., Chen, Y., Jia, L.: Edge-aware U-net with gated convolution for retinal vessel segmentation. *Biomed. Signal Process. Control* **73**, 103472 (2022)
7. Toptacs, B., Hanbay, D.: Retinal blood vessel segmentation using pixel-based feature vector. *Biomed. Signal Process. Control* **70**, 103053 (2021)
8. Tchinda, B.S., Tchiotsop, D., Noubom, M., Louis-Dorr, V., Wolf, D.: Retinal blood vessels segmentation using classical edge detection filters and the neural network. *Inform. Med. Unlocked* **23**, 100521 (2021)
9. Yang, J., Lou, C., Fu, J., Feng, C.: Vessel segmentation using multiscale vessel enhancement and a region based level set mode. *Comput. Med. Imaging Graph.* **85**, 101783 (2020)
10. Dou, Q., Zhang, J., Jiang, P., Tang, H.: Retinal vessel segmentation based on convolutional neural network and connection domain detection. *Procedia Comput. Sci.* **187**, 246–251 (2021)
11. Mayya, V., Sowmya Kamath, S., Kulkarni, U.: Automated microaneurysms detection for early diagnosis of diabetic retinopathy: a comprehensive review. *Comput. Methods Programs Biomed.* **Update** **1**, 100013 (2021)
12. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Karel Zuiderveld, pp. 474–485 (1994)
13. Garcia-Lamont, F., Cervantes, J., López, A., Rodriguez, L.: Segmentation of images by color features: a survey. *Neurocomputing* **292**, 1–27 (2018)
14. Otsu, N.: A threshold selection method from Gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979)
15. Strisciuglio, N., Vento, M., Petkov, N., Azzopardi, G.: *Med. Image Anal.* **19**(1), 46–47 (2015)
16. Fraz, M.M., et al.: An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.* **59**(9), 2538–2548 (2012)

17. Vega, R., Sanchez-Ante, G., Falcon-Morales, L.E., Sossa, H., Guevara, E.: Retinal vessel extraction using Lattice Neural Networks with dendritic processing. *Comput. Biol. Med.* **58**, 20–30 (2015)
18. Liskowski, P., Krawiec, K.: Segmenting retinal blood vessels with deep neural network. *IEEE Trans. Med. Imaging* **35**(11), 2369–2380 (2016)
19. Yan, Z., Yang, X., Cheng, K.-T.: Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. *IEEE Trans. Biomed. Eng.* **65**(9), 1912–1923 (2018)
20. Li, Q., Feng, B., Xie, L., Liang, P., Zhang, H., Wang, T.: A cross-modality learning approach for vessel segmentation in retinal images. *IEEE Trans. Med. Imaging* **35**(1), 109–118 (2016)



Graph-Based Anomaly Detection via Attention Mechanism

Yangming Yu^(✉), Zhiyong Zha, Bo Jin, Geng Wu, and Chenxi Dong

Information and Communication Branch of Hubei Epc, Hubei, China
24870875@qq.com

Abstract. Graph-based anomaly detection aims to spot outliers and anomalies from big data, with numerous high-impact applications in areas such as security, industry, and data auditing. Deep learning-based methods could implicitly identify patterns from data. Recently, graph representation learning based on Deep Neural Network (DNN) have made significant progress. How to use deep learning methods to detect graph anomalies and assist audit work has received extensive attention from academia and industry. Related works have explored the graph-based anomaly detection, but they lack attention to DNN and auditing techniques. This paper investigates the DNN-based graph representation learning and sorts out the graph anomaly detection algorithm based on deep learning and proposes a graph-based anomaly detection algorithm via attention mechanism to assist the audit process.

Keywords: Anomaly detection · Graph learning · Neural network

1 Introduction

As a common data representation structure, graph can be widely used to represent complex structured linked data. Compared with other data structures, it can better represent and store the connections between data content and other entities. In the real world, graphs are widely used in financial big data auditing, industrial structured data analysis, industrial network traffic analysis, Web analysis, traffic road network optimization, knowledge graph construction, and other fields. How to detect the anomalies accurately and rapidly in these large and complex graph data has aroused widespread concern in academia and industry. Graph anomaly detection refers to finding structures (including nodes, edges, etc.) containing uncommon and outlier than common patterns in a giant graph or massive graph representation data, and has a wide range of application scenarios, such as big data auditing, anomalous data finding in industrial networks, malicious attacks in the Internet, breaking news detection in social networks, and spammers discovery in e-commerce etc. Compared with traditional anomaly detection methods, graph-based anomaly detection can visually present complex data and incorporate the implicit correlation in the data into the anomaly detection process due to the powerful representation capability of graph structure.

2 Related Work

2.1 Traditional Graph Anomaly Detection Algorithm

The first work on graph-oriented anomaly detection was published in 2003 [1], and the existing work can be roughly classified into static graph-based method and dynamic graph-based method. In static graph-based anomaly detection work, one class of methods uses ego networks [2] or group-based method [3]; another class of methods performs anomaly detection based on the structural information of the graph [4–6], while some work is based on subspace selection, trying to find anomalies in the subspace of node features [7]. There are also some works using probabilistic and statistical methods to obtain statistical information of graphs for anomaly detection [8–10]. Although these works have made great progress in anomaly detection, these methods, such as those using ego networks, are difficult to achieve good results when the graph is sparse because the interaction between nodes must be considered in processing graph data; or, for example, subspace selection and statistical methods, are difficult to make comprehensive use of node attributes and structure information due to shallow learning mechanisms. In terms of dynamic graph-based anomaly detection, there are also similarly several works based on groups [11–13], on structure [6, 16], or on probabilistic statistics [12, 14] for anomaly detection. Another typical method is to first obtain the outline of the graph and then identify the anomalies in the outline through clustering and anomaly detection, e.g., [15]. However, the outline obtained by these methods cannot retain important structural information, such as the information of adjacent nodes. Most of the existing dynamic graph-based anomaly detection methods rely on heuristic rules and usually simply consider a certain class of features. Although some methods take content and even time factors into consideration, they are not flexible, leading to their applications being limited to specific scenarios.

2.2 Deep Learning Based Graph Anomaly Detection Algorithm

In recent years, deep learning has become an extremely important part of artificial intelligence and machine learning, showing superior performance in extracting potentially complex patterns in data, and has been widely used in areas such as audio, image, and natural language processing. Deep learning methods can reasonably deal with complex attribute information and can learn implicit patterns from data. Moreover, the embedding of graphs through neural networks can not only retain information well [17], but also process the attributes of nodes or edges well while preserving structural information, thus facilitating the inspection of the similarity of nodes or edges in hidden space. With the remarkable progress of embedding representations of graphs in recent years, how to use deep learning methods for graph anomaly detection has attracted a lot of attention in the past few years. Deep learning-based graph anomaly detection methods usually use the embedding representation of the graph to first represent the graph as a vector in the hidden space, then use the vector to reconstruct the graph so as to remove the effect of anomalous information, and finally perform anomaly detection by reconstruction error.

There have been very comprehensive review articles about anomaly and outlier detection. For example, McConville et al. [18] focus on high-dimensional outlier detection,

and Mikolov et al. [19] discuss local outlier detection techniques. However, these articles generally focus on the point of multidimensional data instances and do not or not directly focus on graph-based detection techniques. Although [20] investigates anomaly detection techniques from perspective of graphs, it lacks a focus on graph anomaly detection techniques based on deep learning techniques. Different from previous reviews on anomaly detection, this paper focuses on anomaly detection in large graph or massive graph databases, and comprehensively composes and summarizes deep learning-based graph anomaly detection techniques. It is the earliest research survey focusing on deep learning-based graph anomaly detection techniques.

In this paper, we first make a comprehensive analysis of the definition of anomalies on graphs, and then introduce the graph representation learning methods based on deep neural networks in detail. Then, from the perspectives of static and dynamic graphs, the existing graph anomaly detection methods based on deep learning are systematically summarized and classified, and the limitations of relevant methods are discussed. After that, the practical application scenarios and relevant datasets of graph anomaly detection technology are briefly introduced. Finally, we discussed the challenges and feasible future research directions of deep learning-based graph anomaly detection. In this paper, we expect to provide the ideas that can be drawn from the current research status of deep learning-based graph anomaly detection for the subsequent research.

3 Graph-Based Anomaly Detection via Attention Mechanism

The key problem of anomaly detection is how to extract feature from structured graph, embedding graph data to a low dimensional space. It has become the hot spot of industry and academia that how graph representation learning can support inference and maintain the graph structure at the same time. Since encoding representation of nodes could be seen as the input of anomaly detection algorithm, graph representation learning has important impact in the field of accounting or auditing.

In this section, we firstly introduce the representation of structural instance feature via graph-based attention mechanism. Secondly, we improve the traditional anomaly detection methods from using the optimal transmission scheme of single sample and standard sample mean to learn the outlier probability. And we further detect anomaly attributes from some samples with the guidance of this probability.

3.1 Graph Representation Learning Based on Deep Neural Network

The pipeline of anomaly detection network based on graph encoding was shown in Fig. 1:

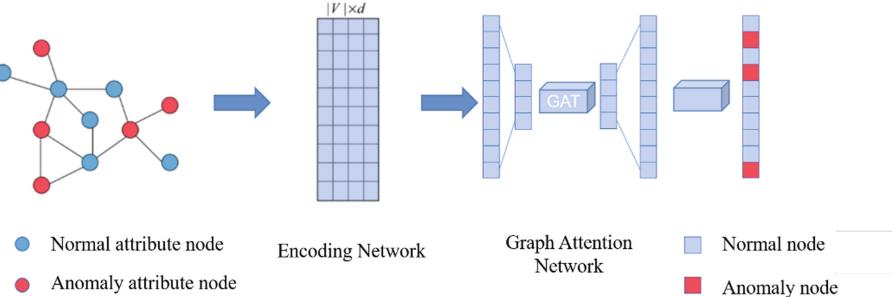


Fig. 1. The framework of our anomaly detection network.

The Encoding Representation of Structural Data. The structural data should be projected to some latent variables. This allows convenient process for post networks. As shown in Fig. 2, We design an auto-encoder framework to compress structural data.

This framework helps to encode the input data. And we use the L2 distribution of input and output data as loss to supervise network training. We use it to project input data into desired dimension. Meanwhile, the loss of auto-encoder can be added in end-to-end network training as an auxiliary loss, which leads to a better result.

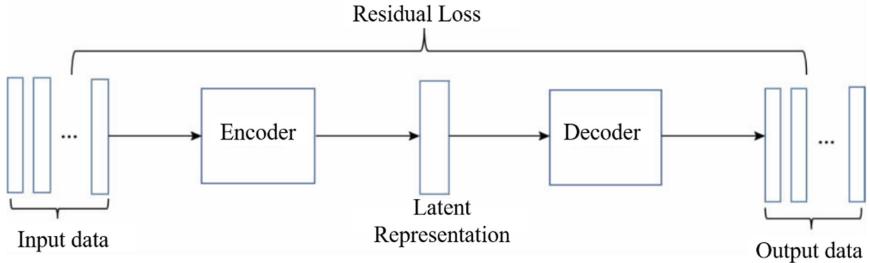


Fig. 2. The framework of auto-encoder.

Latent Feature Extraction, and Local-Global Feature Fusion via Attention Network. To get more robustness and discernible representation, we fuse local and global feature extracted from auto-encoder. The structural data D is fed into coarse feature model M to get the latent representation F.

We use multi-layer Graph Neural Network (GNN) to fuse feature. The category encoder E_c was designed to encode different data category into category encoding feature f_{cenc} . And we fuse it with data encoding feature f to get f_{mix} :

$$f_{cenc} = E_c(f)$$

$$f_{mix} = f_{kenc} + f$$

Then we use multi-layer GNN to further aggregate the fused feature. Specifically, the feature O_l output from l layer of GNN is than aggregate by a self-attention module.

$$O_0 = f_{mix}$$

$$attn_output(O_l) = \text{Attention}(M_q O_l, M_k O_l, M_v O_l)$$

$$\text{Attention } (Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

M_q, M_k, M_v are the query parameter matrices and a set of key-value pairs parameter matrices. And d_k represent the feature dimension. Then, the output $attn_output(O_l)$ of GNN is processed by a Multi-layer Perceptron (MLP):

$$SubLayer(O_l) = \text{MLP}(attn_output(O_l))$$

O_{l+1} was calculated as below:

$$O_{l+1} = O_l + SubLayer(O_l)$$

We design a differentiable pooling matrix to fuse the feature from original graph and subgraph. This network could fuse feature matrix F from original graph and $F_{subgraph}$ from subgraph:

$$F_{subgraph} = M_{assignment} * F$$

where the matrices $O \in \mathbb{R}^{n \times d}, F_{subgraph} \in \mathbb{R}^{m \times d}$ and $M_{assignment} \in \mathbb{R}^{n \times m}$. n, m represents the number of nodes in original graph and subgraph respectively, and d represents the feature dimension of node. $M_{assignment}$ is a feature assignment matrix, which used to re-project m subgraph features to n graph features. Lastly, the mix feature O_{mix} can be calculated as below:

$$O_{mix} = {M_{assignment}}^T * O_{subgraph} + O$$

Anomaly Detection Classifier Based on Optimal Transport. After setting the threshold of anomaly node, the anomaly detection can be seen as an optimal transport problem to get the allocation scheme from initial status to target status. And we use Sinkhorn algorithm to solve it, which build the similarity degree matrix from all data and transport cost matrix C to represent the relevance between attribute feature i and j of instance and mean instance respectively:

$$c_{ij} = -\log(S_{ij})$$

where c_{ij} represents the relationship between attribute feature i and j of instance and the mean instance of all samples respectively. The column i and row j of the maximum value of matrix S represents the similarities between them.

Loss Function. Anomaly detection is a binary classification problem. We could use cost c to train our model. The loss function can be formulated as below:

$$L = - \sum_{i \in N, j \in M} \log p_{i,j}$$

where N represents the number of instances in a batch and M represents the number of instances in referenced standard data.

4 Experiments

4.1 Static Graph Anomaly Detection Dataset

In order to verify the effectiveness of the above deep learning-based anomaly detection methods for static graphs, this section will use the codes of two published sources [21, 22] to conduct comparative experiments on public datasets to evaluate their performance. The basic information of the datasets used is shown in Table 1. Since the existing public datasets usually have no anomaly annotations, we manually inject 5% of the anomalies (including structural anomalies, attribute anomalies, and anomalies composed of the two) into a partially public attribution network dataset. The strategy used in [22] is followed to ensure that the injected exception is close to the actual exception.

Table 1. Experimental data information.

Dataset	Number of labels	Dimension of attributes	Normal/anomaly Numbers
WebKB	5	1703	877/919
Cora	7	1433	2708/2843
Citeseer	6	3703	3312/3477

Based on the above three datasets, we tried the performance of existing methods and the method proposed in this paper under different recall percentage thresholds L%.

4.2 Experimental Results of Static Image Anomaly Detection Dataset

It can be seen from Table 2 that the recall of our scheme for abnormal samples under different percentage thresholds on the three static data sets is higher than that of other comparison schemes. On the Citeseer dataset, the recall is more than 7% higher than that of the better performing Dominant algorithm.

Table 2. Static graph anomaly detection results.

Dataset	Method	Recall@L = 5	Recall@L = 10	Recall@L = 20
WebKB	Dominant	0.09	0.13	0.21
	Done	0.05	0.12	0.18
	Ours	0.11	0.14	0.22
Cora	Dominant	0.64	0.71	0.74
	Done	0.03	0.08	0.19
	Ours	0.65	0.73	0.79
Citeseer	Dominant	0.33	0.41	0.52
	Done	0.31	0.35	0.52
	Ours	0.41	0.49	0.59

4.3 Anomaly Detection Dataset Based on Power Grid Data

There is a variety of unstructured data in the power grid. For example, according to the source of data, there are firewall logs, intrusion detection system logs, and logs generated by business systems; according to the types of logs, there are traffic logs, configuration management logs, and security attack logs. These logs are designed for specific needs, they record different network information, user behavior and system operations. We structured these data and formed corresponding datasets. By artificially modifying these datasets, datasets with real anomaly labels are obtained. At the same time, we also use existing open source datasets for sampling and construction to adapt the network input in a unified structured way. We conduct experiments on this real power grid dataset to verify the effectiveness of the algorithm (Table 3).

Table 3. Construction scheme of fixed assets card data set based on power grid data set

Asset exception	Feature description	Eigenvalue selection	Number of normal/abnormal constructs
Asset categorization exception	Asset description does not match asset class	Asset class, asset description, quantity, unit of measure	900/100

(continued)

Table 3. (*continued*)

Asset exception	Feature description	Eigenvalue selection	Number of normal/abnormal constructs
data integrity exception	Key asset attribute data is null	Asset description, quantity, unit of measurement, original value of asset, accumulated depreciation, depreciation method, useful life	1800/200
Asset management exception	Administrative requirements not implemented	Asset class, custodian used, quantity, how assets change	900/100
Asset Data Abnormal	Administrative requirements not implemented	Asset class, original value of assets, net book value, accumulated depreciation, voltage level	900/100

After comprehensive analysis, 12 characteristic values are determined: asset code, asset type, asset change method, asset status, measurement unit, quantity, voltage level, expected service life, use custodian, original value of assets (initial acquisition value), net book value, Accumulated depreciation amount. The characteristic data is divided into two categories: discrete data and continuous data, which are divided as follows: (1) Discrete variables, also known as text variables, classification variables or enumeration variables, present a discrete state. Including asset class, asset description, asset change method, asset status, measurement unit, voltage level, use custodian; (2) Continuous variable: can take any value within a certain range, and the value is continuous, including quantity, Estimated useful life, original value of assets, net book value, and accumulated depreciation.

4.4 Experimental Results

After completing the above construction, we adapt the input format so that the dataset can be directly connected to the three baseline methods for comparison. The results are shown in the following table (Table 4):

Table 4. Experimental results of anomaly detection in power grid datasets.

Method	Recall@10			
	Asset exception	Data integrity exception	Asset management exception	Asset Data Abnormal
(a) Logistics	0.57	0.83	0.67	0.79
(b) Dominant	0.66	0.89	0.88	0.83
(c) Done	0.59	0.81	0.78	0.77
(e) Ours	0.75	0.93	0.89	0.84

Our proposed graph network-based anomaly data detection scheme achieves the best results on all four anomalies. Among these methods, Logistics is the simplest logistic regression scheme. It can be seen that its effect on data integrity anomalies and asset data anomalies is good, and the effect is relatively close to our scheme. However, in the classification task of asset anomaly and asset management anomaly, due to the discrepancy of attribute discretization, the classification effect is not as good as that based on continuous variables. Furthermore, we count the operating efficiency of these methods. Among them, the logistic regression scheme occupies the least memory and running speed. Our method occupies more memory, but the speed is faster than the previous better Dominant scheme under the premise of obtaining optimal results (Table 5).

Table 5. Recognition accuracy and recognition time of three kinds of test data

Method	Memory (MB)	Time(s)	Method
Logistics	15	0.082	Logistics
(b) Dominant	105	1.112	(b) Dominant
(c) Done	75	0.099	(c) Done
(e) Ours	375	0.752	(e) Ours

5 Conclusion

This paper uses graph attention neural network to conduct experiments on traditional static graph anomaly detection and data based on real power grid data and proposes an anomaly detection scheme based on multi-graph attention. Finally, the effectiveness of the proposed method is verified by experiments. --Better results are achieved on both traditional static graphs and power grid data. With the continuous advancement of informatization, more and more industrial applications have added machine learning auditing and anomaly detection mechanisms. In the future, we will continue to carry out further research on the audit scheme based on graph neural network.

References

1. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 679–698 (1986)
2. Noble, C.C., Cook, D.J.: Graph-based anomaly detection. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 631–636 (2003)
3. Perozzi, B., Akoglu, L.: Scalable anomaly ranking of attributed neighborhoods. In: Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 207–215 (2016)
4. Aggarwal, C.C.: An introduction to outlier analysis. In: Outlier Analysis, pp. 1–34. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-47578-3_1
5. Colladon, A.F., Remondi, E.: Using social network analysis to prevent money laundering. *Expert Syst. Appl.* **67**, 49–58 (2017)
6. Manjunatha, H.C., Mohanasundaram, R.: BRNADS: big data real-time node anomaly detection in social networks. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 929–932. IEEE (2018)
7. Perozzi, B., Akoglu, L., Iglesias Sánchez, P., et al.: Focused clustering and outlier detection in large attributed graphs. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1346–1355 (2014)
8. Dai, H., Zhu, F., Lim, E.P., et al.: Detecting anomalies in bipartite graphs with mutual dependency principles. In: 2012 IEEE 12th International Conference on Data Mining, pp. 171–180. IEEE (2012)
9. Sánchez, P.I., Müller, E., Irmler, O., et al.: Local context selection for outlier ranking in graphs with multiple numeric node attributes. In: Proceedings of the 26th International Conference on Scientific and Statistical Database Management, pp. 1–12 (2014)
10. Tsang, S., Koh, Y.S., Dobbie, G., et al.: SPAN: finding collaborative frauds in online auctions. *Knowl.-Based Syst.* **71**, 389–408 (2014)
11. Shehnepoor, S., Salehi, M., Farahbakhsh, R., et al.: NetSpam: a network-based spam detection framework for reviews in online social media. *IEEE Trans. Inf. Forensics Secur.* **12**(7), 1585–1595 (2017)
12. Carvalho, L.F.M., Teixeira, C.H.C., Meira, W., et al.: Provider-consumer anomaly detection for healthcare systems. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI), pp. 229–238. IEEE (2017)
13. Giatsoglou, M., Chatzakou, D., Shah, N., Faloutsos, C., Vakali, A.: Retweeting activity on twitter: Signs of deception. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS (LNAI), vol. 9077, pp. 122–134. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18038-0_10
14. Hooi, B., Song, H.A., Beutel, A., et al.: Fraudar: bounding graph fraud in the face of camouflage. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 895–904 (2016)
15. Dang, Q., Zhou, Y., Gao, F., et al.: Detecting cooperative and organized spammer groups in micro-blogging community. *Data Min. Knowl. Disc.* **31**(3), 573–605 (2017)
16. Bhattacharjee, S.D., Yuan, J., Jiaqi, Z., et al.: Context-aware graph-based analysis for detecting anomalous activities. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 1021–1026. IEEE (2017)
17. Ranshous, S., Shen, S., Koutra, D., et al.: Anomaly detection in dynamic networks: a survey. *Wiley Interdisc. Rev. Comput. Stat.* **7**(3), 223–247 (2015)
18. McConville, R., Liu, W., Miller, P.: Vertex clustering of augmented graph streams. In: Proceedings of the 2015 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 109–117 (2015)

19. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
20. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Min. Knowl. Disc. **29**(3), 626–688 (2015)
21. Ding, K., Li, J., Bhanushali, R., et al.: Deep anomaly detection on attributed networks. In: Proceedings of the 2019 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 594–602 (2019)
22. Bandyopadhyay, S., Vivek, S.V., Murty, M.N.: Outlier resistant unsupervised deep architectures for attributed network embedding. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 25–33 (2020)



A Classification Algorithm Based on Discriminative Transfer Feature Learning for Early Diagnosis of Alzheimer's Disease

Xinchun Cui^{1(✉)}, Yonglin Liu¹, Jianzong Du², Qinghua Sheng³, Xiangwei Zheng⁴, Yue Feng⁵, Liying Zhuang⁶, Xiuming Cui¹, Jing Wang¹, and Xiaoli Liu^{6(✉)}

¹ School of Computer Science, Qufu Normal University, Rizhao 276826, China
cxcscsd@126.com

² Department of Respiratory Medicine, Zhejiang Hospital, Hangzhou 310013, China

³ Pharmacy Department of Rizhao Central Hospital, Rizhao 276800, China

⁴ School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

⁵ Department of Radiology, Zhejiang Hospital, Hangzhou 310013, China

⁶ Department of Neurology, Zhejiang Hospital, Hangzhou 310013, China

Abstract. This paper proposes a discriminative transfer feature learning method for MCI conversion prediction using data from the target domain and the auxiliary domain. A transfer component analysis method based on the Maximum Mean Discrepancy (MMD) is proposed at first, which is used to weaken the difference of data distribution between the relevant domain and the target domain. Next, the discriminant optimization term is added to measure the correlation between the sample categories and the sample features of the auxiliary domain, and to improve the inter-class separability of the algorithm. Finally, the support vector machine (SVM) is used to classify MCI patients.

Keywords: AD · Mild cognitive impairment · Transfer learning · SVM

1 Introduction

Alzheimer's disease (AD) is the most common type of dementia, comprising an estimated 60–80% of all dementia cases, which is an irreversible neurodegenerative disease [1–9]. Therefore, considering the good performance of transfer learning in domain learning, it has been successfully introduced into medical image analysis [10–14]. In this paper, we propose a discriminative transfer feature learning (DTFL) algorithm, which combines discriminative optimization term and transfer component analysis method to extract a subset of relevant discriminant features for MCI conversion prediction.

2 Materials

Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The primary goal of ADNI has been

to test whether serial Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

In order to evaluate the effectiveness of the proposed method, we choose the ADNI database (<http://adni.loni.usc.edu/>) T1-weighted MRI mode data of the subjects for experiments, and only select the benchmark time of these subjects to collect MRI images. For the baseline diagnostic classification, we used a total of 197 subjects scanned from a 1.5T scanner. As shown in Table 1, demographic information of the 197 subjects and information on MRI image acquisition are presented in the references [15].

Table 1. Statistical information of subjects (mean standard deviation)

Diagnosis	Subjects	Age	Gender (F/M)	MMSE	CDR
AD	51	75.8 ± 7.5	23/28	23.6 ± 2.2	0.7 ± 0.3
NC	50	77.8 ± 6.8	27/23	28.8 ± 1.4	0.0 ± 0.0
MCIc	51	72.5 ± 6.5	26/25	26.7 ± 1.3	0.5 ± 0.0
MCInc	45	71.9 ± 7.6	20/25	27.3 ± 1.6	0.5 ± 0.0

CDR: clinical dementia rating scale, 0 = no dementia, 0.5 = suspected dementia, 1 = mild dementia, MMSE: Concise mental state examination scale

3 Method

In this section, we first briefly introduce our proposed feature learning method, and then we introduce the MR image preprocessing and feature extraction methods; Finally, we introduce the diagnostic model of discriminative transfer feature learning (DTFL) and the corresponding optimization methods.

3.1 Mainframe of Proposed Approach

Our method consists of three main parts: (1) image preprocessing and feature extraction; (2) discriminative transfer feature learning (DTFL); and (3) brain disease classification using SVM. The specific process is as follows. We first preprocess all MR images and extract features from the MR images. Then, we use the proposed discriminative transfer feature learning (DTFL) method to select information features. We eventually train the SVM classifier to diagnose and predict MCI using dimensionality reduction data in the target domain. This is illustrated in Fig. 1.

3.2 Image Preprocessing and Feature Extraction

The original MRI image acquired requires a series of image pre-processing and ROI-based features extraction, which are input to the classifier for classification. We first perform an anterior commissure (AC)-posterior commissure (PC) correction on all MRI

and PET images using MIPAV software. The AC-PC corrected image is resampled to $256 \times 256 \times 256$, and N3 algorithm is used to correct the intensity inhomogeneity. For MRI images, skull dissection is performed using VBM software, and the results of skull dissection are manually examined to ensure removal of the skull and dura mater. Then, the image of skull dissection is registered to a manually labeled cerebellar template. All voxels are removed from the labeled cerebellar mask, and the cerebellum is removed. Finally each image is divided into three distinct tissues using VBM software package: grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). As the atlas warped, each subject is registered on an AAL template with 90 automatically marked regions of interest (ROIs). The volume of the GM tissue in the ROI is calculated as a feature for each ROI. Each participant extracted 90 MRI features.

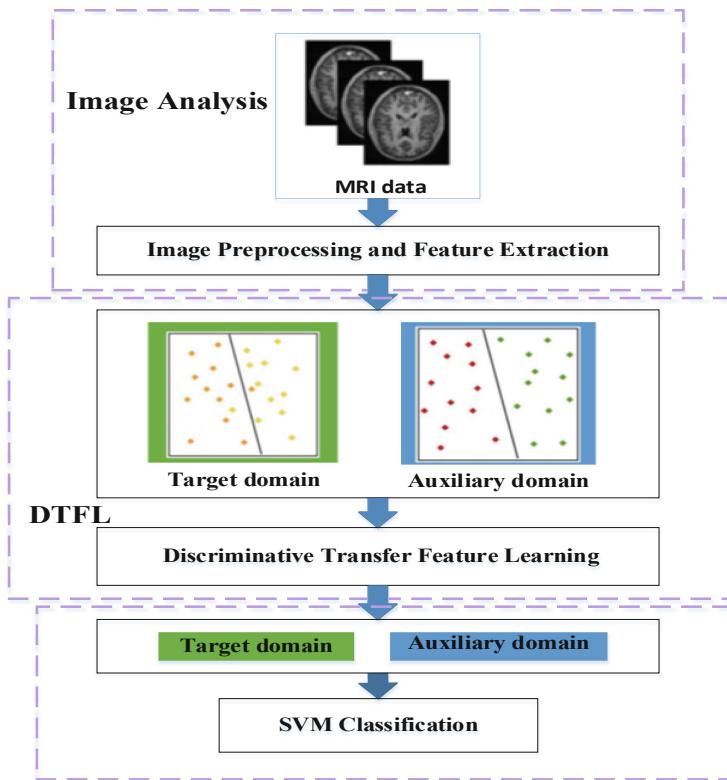


Fig. 1. Illustration of our proposed framework for Alzheimer's disease prediction

3.3 Transfer Component Analysis

Transfer Component Analysis (TCA) [15] is a transfer learning method based on feature mapping, which is used for transfer learning of samples with same feature space and

different edge distribution. The core idea of TCA algorithm is to map the features of two sample sets to the same hidden space by setting a unified kernel function. To minimize the data distribution distance between domains, the Maximum Mean Discrepancy (MMD) are used as the measurement criteria.

We assume the source domain data as $D_S = \{(x_{S_1}, y_{S_1}), \dots, (x_{S_{n_1}}, y_{S_{n_1}})\}$, where $x_{S_i} \in X$ and $y_{S_i} \in Y$ are input and output respectively. Similarly, we denote the target domain data as $D_T = \{(x_{T_1}, y_{T_1}), \dots, (x_{T_{n_2}}, y_{T_{n_2}})\}$, the input is $x_{T_i} \in X$ and the output is unknown. In addition, x_{S_i} and x_{T_i} correspond to the original space X , and their marginal distributions are $P(X_S)$ and $Q(X_T)$ respectively. In fact, P and Q are completely different to some extent. Considering the noise and the instability of observations, the transfer component analysis technique breaks the hypothesis.

Then a universal kernel transformation Φ is introduced, through which the original space X can be spanned to the subspace H , and all data can be mapped to the subspace by kernel Φ technique. Therefore, the marginal distribution and conditional distribution are redefined as $P(\Phi(X_S)) \approx P(\Phi(X_T))$ and $P(Y_S|\Phi(X_S)) = P(Y_T|\Phi(X_T))$, respectively.

TCA uses a classic distance: Maximum Mean Discrepancy (MMD) to define the distance between two distributions P and Q as follows:

$$Dist(X_S, X_T) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_{S_i}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_{T_i}) \right\|_H^2 \quad (1)$$

where H is a general RKHS, and Φ is a mapping function $X-H$.

According to the advantages of general kernel mapping Φ , the distance between P and Q can be expressed as:

$$Dist(X'_S, X'_T) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_{S_i}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_{T_i}) \right\|_H^2 = tr(KL) \quad (2)$$

TCA introduces a kernel matrix K and L, as follows:

$$K = \begin{pmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{pmatrix} \in \mathbb{R}^{(n_s+n_t) \times (n_s+n_t)} \quad (3)$$

$$L_{ij} = \begin{cases} \frac{1}{n_2}, & x_i, x_j \in X_S \\ \frac{1}{n_2}, & x_i, x_j \in X_T \\ -\frac{1}{n_1 n_2}, & otherwise \end{cases} \quad (4)$$

where $K_{S,S}$, $K_{T,T}$, $K_{S,T}$, $K_{T,S}$ is the kernel matrix obtained from the source domain, the target domain and the mixed domain respectively.

3.4 Discriminative Transfer Feature Learning (DTFL)

The TCA algorithm only considers the correlation between the samples of the auxiliary domain and the target domain, and ignores the local differences between the samples of each domain, such as the compactness of the intra-class samples and the separability of the inter class samples. This may make the original similar data in the projection after a

large deviation. Therefore, linear discriminant analysis (LDA) is introduced to solve the compactness of intra-class samples and the separability of inter-class samples in source domain and target domain data, and improve the correlation between sample categories and sample features.

3.5 Classification

In this paper, we adopt the classification method based on Support Vector Machine (SVM) [16]. The SVM classifier is implemented in the LIBSVM toolbox [17]. The optimal parameters in SVM are obtained in the training set through ten-fold cross-validation.

4 Experiment and Results

This method is compared with several different classification methods, namely SVM, Lap_SVM, TCA_SVM, DTSVM. It is worth noting that the same training and test data are used in all methods for fair comparison. The performance of each comparison method was evaluated by classifying MCIC and MCInc subjects. The ten-fold cross-validation strategy was used to evaluate the classification effect.

In order to evaluate the effectiveness of the proposed method, the classification performance of MCIC and MCInc patients was evaluated by comparing with several different classification methods. Here, “DTSVM” represents a domain transfer support vector machine method for transfer cross-domain kernel learning of auxiliary domain knowledge [10]. Note that each value in Table 2 is the average of the cross-validations performed 10 different times. In addition, we plot the ROC curves achieved by different methods in Fig. 2.

Table 2. Classification performance of MRI samples (%)

Method	ACC (%)	SEN (%)	SPE (%)	AUC
SVM	58.33	55.10	61.70	0.6157
Lap_SVM	61.46	59.52	62.96	0.6571
TCA_SVM	65.54	63.87	69.00	0.6875
DTSVM	73.40	74.30	72.10	0.7266
Proposed	75.13	75.60	76.61	0.7957

In order to further show the superiority of our proposed method, we compare the classification performance of the proposed method with other methods. Table 3 lists the classification performance of different methods. From Table 3, we can observe that our proposed method is superior to the comparison method in classification accuracy and sensitivity in the classification of MCIC and MCInc patients. When comparing with Cheng et al.’s results, the results are competitive. our method improves accuracy by

1.73%. However, we would like to emphasize the improvements of 1.73% in sensitivity, which is clinically regarded more important than other metrics. High sensitivity may be helpful in convincing AD diagnosis, which may be useful in clinical practice.

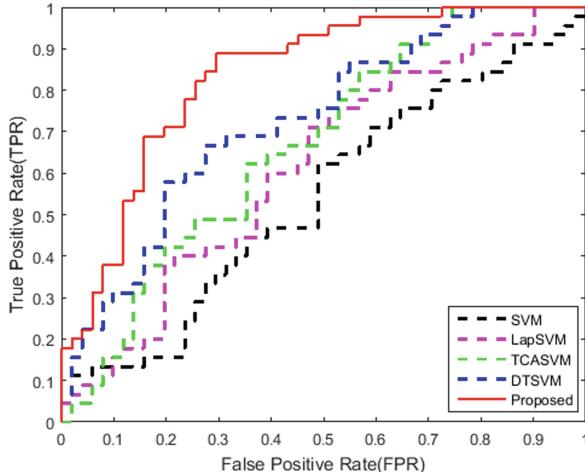


Fig. 2. ROC curve and AUC values achieved by different methods

Table 3. Comparison with other methods

Method	Feature	Classifier	Subjects	ACC (%)	SEN (%)	SPE (%)
Cuingn [18]	Voxel-Stand-D GM	SVM	76MCIc+134MCInc	70.40	57.00	78.00
Zhang [6]	ROI GM	SVM	43MCIc+48MCInc	62.00	56.60	60.00
Cheng [10]	ROI GM	DTSVM	51AD+52NC 43MCIc+56MCInc	73.40	74.30	72.10
Proposed	ROI GM	TCA-LDA	51AD+50NC 45MCIc+51MCInc	75.13	75.60	76.61

5 Conclusion

We proposed a classification framework based on transfer feature learning. According to the correlation between the target domain (MCI) and the auxiliary domain (AD and NC) data, we propose a DTFL algorithm to effectively combine the discriminative optimization term and the transfer component analysis method to extract the most relevant correlation. A subset with discriminative features is used for MCI diagnosis prediction. Our experiments on a subset of ADNI dataset show that the proposed method is effective for MCI prediction and diagnosis.

References

1. Bain, L.J., Jedrzejewski, K., Morrison-Bogorad, M., et al.: Healthy brain aging: a meeting report from the Sylvan M. Cohen annual retreat of the University of Pennsylvania Institute on Aging. *Alzheimers Dement.* **4**(6), 443–446 (2008)
2. Hinrichs, C., Singh, V., Xu, G., et al.: Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* **55**(2), 574–589 (2011)
3. Querbes, O., Aubry, F., Pariente, J., et al.: Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain* **132**(8), 2036–2047 (2009)
4. Aksu, Y., Miller, D.J., Kesidis, G., et al.: An MRI-derived definition of MCI-to-AD conversion for long-term, automatic prognosis of MCI patients. *PLoS ONE* **6**(10), e25074 (2011)
5. Cho, Y., Seong, J.K., Jeong, Y., et al.: Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *Neuroimage* **59**(3), 2217–2230 (2012)
6. Zhang, D., Shen, D., Alzheimer's Disease Neuroimaging Initiative: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage* **59**(2), 895–907 (2012)
7. Li, H., Liu, Y., Gong, P., et al.: Hierarchical interactions model for predicting Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) conversion. *PLoS ONE* **9**(1), e82450 (2014)
8. Zhu, X., Suk, H.I., Wang, L., et al.: A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Med. Image Anal.* **38**, 205–214 (2017)
9. Lei, B., Chen, S., Ni, D., et al.: Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. *Front. Aging Neurosci.* **8**, 77 (2016)
10. Cheng, B., Liu, M., Zhang, D., et al.: Domain transfer learning for MCI conversion prediction. *IEEE Trans. Biomed. Eng.* **62**(7), 1805–1817 (2015)
11. Wimmer, G., Vécsei, A., Uhl, A.: CNN transfer learning for the automated diagnosis of celiac disease. In: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications, pp. 1–6. IEEE (2016)
12. Cheng, B., Liu, M., Suk, H.-I., Shen, D., Zhang, D.: Multimodal manifold-regularized transfer learning for MCI conversion prediction. *Brain Imaging Behav.* **9**(4), 913–926 (2015). <https://doi.org/10.1007/s11682-015-9356-x>
13. Colbaugh, R., Glass, K., Gallegos, G.: Ensemble transfer learning for Alzheimer's disease diagnosis. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3102–3105. IEEE (2017)
14. Cheng, B., Liu, M., Shen, D., Li, Z., Zhang, D.: Multi-domain transfer learning for early diagnosis of Alzheimer's disease. *Neuroinformatics* **15**(2), 115–132 (2017). <https://doi.org/10.1007/s12021-016-9318-5>
15. Zhang, D., Wang, Y., Zhou, L., et al.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55**(3), 856–867 (2011)
16. Pan, S.J., Tsang, I.W., Kwok, J.T., et al.: Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **22**(2), 199–210 (2010)
17. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 1–27 (2011)
18. Cuingnet, R., Gerardin, E., Tessieras, J., et al.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* **56**(2), 766–781 (2011)

19. Eskildsen, S.F., Coupé, P., García-Lorenzo, D., et al.: Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage* **65**, 511–521 (2013)
20. Min, R., Wu, G., Cheng, J., et al.: Multi-atlas based representations for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* **35**(10), 5052–5070 (2014)
21. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7**(11), 2399–2434 (2006)



A Systematic Study for the Effects of PCA on Hyperspectral Imagery Denoising

Guang Yi Chen¹(✉) and Wen Fang Xie²

¹ Department of Computer Science and Software Engineering, Concordia University, Montreal,
QC H3G 1M8, Canada

guangyi_chen@hotmail.com

² Department of Mechanical and Industrial Engineering, Concordia University, Montreal,
QC H3G 1M8, Canada

wfxie@me.concordia.ca

Abstract. In this paper, we conduct a study for the effects of principal component analysis (PCA) on hyperspectral imagery denoising. Our previous works combined PCA with wavelet shrinkage, block matching and 3D filtering (BM3D), and block matching and 4D filtering (BM4D), respectively, and very good denoising results have been obtained for hyperspectral imagery with very little noise. To demonstrate if these methods are the best for other noise scenarios, we combine PCA with video BM3D (VBM3D) and video BM4D (VBM4D), and non-local means (NL-Means) as well. Experimental results show that the PCA+VBM3D is the best denoising method for moderate and high noise levels, and PCA+BM4D is preferred for very low noise levels.

Keywords: Hyperspectral imagery denoising · BM3D · BM4D · VBM3D · VBM4D · Principal component analysis (PCA)

1 Introduction

Reducing the noise in hyperspectral imagery data cubes is a very important preprocessing step in many remote sensing applications. For example, it is useful in object classification, endmember extraction, target detection, mineral detection, environment monitoring, military surveillance, and so forth. In today's hyperspectral imagery data cubes, only a very small amount of noise is present in them, and the noise can hardly be seen by human's eyes. However, reducing this small amount of noise is still desirable in remote sensing applications.

Many methods have been developed for hyperspectral imagery denoising in recent years. Ye et al. [1] proposed a multi-task sparse nonnegative matrix factorization for joint spectral-spatial hyperspectral imagery denoising. Yuan et al. [2] studied hyperspectral image denoising by employing a spectral-spatial adaptive total variation model. Othman and Qian [3] worked on noise reduction of hyperspectral imagery by using hybrid spatial-spectral derivative-domain wavelet shrinkage. Zhao and Yang [4] investigated hyperspectral image denoising by means of sparse representation and low-rank constraint.

In addition to many existing methods for hyperspectral imagery denoising, we also published three papers on this topic in [5, 6] and [7] by using principal component analysis (PCA) to preprocess the data cube. We then only reduce the noise in the noisy PCA components but keep the first few PCA components untouched. An inverse PCA will generate the denoised datacube. Method [5] performs 2D wavelet shrinkage to every PCA component and then a 1D wavelet-based denoisng along the spectral direction. Methods [6] and [7] combines PCA with block matching and 3D filtering (BM3D) and block matching and 4D filtering (BM4D), respectively, for hyperspectral imagery denoising.

We briefly review several denoising methods used in this paper.

- BM3D [8] is currently the state-of-the-arts image denoising method, which proposes a novel image denoising strategy based on an enhanced sparse representation in transform domain. The enhancement of the sparsity is achieved by grouping similar 2D fragments of the image into 3D data arrays. Collaborative filtering is a special procedure developed to deal with these 3D groups. It includes three successive steps: 3D transformation of a group, shrinkage of transform spectrum, and inverse 3D transformation. Thus, the obtained 3D estimation of the group consists of an array of jointly filtered 2D fragments. Due to the similarity between the grouped blocks, the transform can achieve a highly sparse representation of the image so that the noise can be well reduced by shrinkage.
- VBM3D [9] is a video denoising method by means of grouping and collaborative filtering. Grouping is performed by a specially developed predictive-search block matching technique that significantly reduces the computational cost of the search for similar blocks. A two-step video-denoising algorithm where the predictive search block-matching is combined with collaborative hard thresholding in the first step and with collaborative Wiener filtering in the second step. At a reasonable computational cost, this algorithm achieves state-of-the-art denoising results in terms of both PSNR and visual quality.
- BM4D [10] implements the grouping and collaborative filtering paradigm, where mutually similar D-dimensional patches are stacked together in a (D+1)-dimensional array and jointly filtered in the transform domain. Unlike BM3D where the basic data patches are blocks of pixels, the BM4D utilizes cubes of voxels, which are stacked into a four-dimensional group. The 4D transform applied on the group simultaneously exploits the local correlation present among voxels in each cube and the nonlocal correlation between the corresponding voxels of different cubes. Thus, the spectrum of the group is highly sparse, leading to very effective separation of signal and noise through coefficient shrinkage. After inverse transformation, the method obtains estimates of each grouped cube, which are then adaptively aggregated at their original locations.
- In VBM4D [11], groups are 4-D stacks of 3-D volumes, and the collaborative filtering is performed via a separable 4-D spatial-temporal transform. The transform leverages three types of correlation that characterize natural video sequences: local spatial correlation between pixels in each block of a volume, local temporal correlation between blocks of each volume, and nonlocal spatial and temporal correlation between volumes of the same group. The 4-D group spectrum is highly sparse, which

makes the shrinkage more effective than in V-BM3D, yielding superior performance of V-BM4D in terms of noise reduction.

- Non-local means (NL-Means [12]) method is an image denoising method. Local-means filters take the mean value of a group of pixels surrounding a target pixel to smooth the image, whereas non-local means filtering takes a mean of all pixels in the image, weighted by how similar these pixels are to the target pixel. This result is much greater post-filtering clarity, and less loss of detail in the image compared with local mean algorithms.

2 Proposed Study

We have developed PCA+Wavelet shrinkage, PCA+BM3D, and PCA+BM4D for hyperspectral imagery denoising in our previous research. The PCA transform [13] compresses most of the information of an N-dimensional data set in the first a few PCA output components, so it is desirable to reduce the noise in the low-energy output components, which contain most of the noise but not the high-energy output components that contain most of the information. In this way, we can retain most fine features in the hyperspectral data cube. However, for our three denoising methods [5–7], we only considered two hyperspectral imagery data cubes with very little noise. It is not clear whether these methods will perform well for moderate and high noise environment. In this section, we will also study PCA+VBM3D, PCA+VBM4D, and PCA+NL-Means for moderate and high noise. For these noise scenarios, we create noisy data cube by adding Gaussian white noise to the clean data cube as:

$$B = A + s_n Z, \quad (1)$$

where Z obeys normal distribution with zero mean and unit variance, A is the clean data cube, and B is the noisy data cube, and s_n is the noise standard deviation. Here, s_n takes the values of $\{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ in our studies. If we normalize the input hyperspectral imagery data cubes to the range of $[0, 255]$, these s_n correspond to $\{2.81, 5.62, 8.43, 11.24, 14.05, 16.86, 19.67, 22.48, 25.29, 28.10\}$ for the Greater Victoria Watershed District (GVWD) data cube, and $\{3.84, 7.69, 11.53, 15.37, 19.22, 23.06, 26.90, 30.74, 34.59, 38.43\}$ for the Cuprite data cube. We draw the flow chart of the proposed denoising technique by combining PCA with existing denoising methods (see Fig. 1). In the figure, the ‘Denoising Methods’ can be wavelet shrinkage, BM3D, VBM3D, BM4D, VBM4D, or NL-Means.

In [5], we selected the number of PCA components to be denoised manually. This is undesirable in practical remote sensing applications. Therefore, we developed in [7] a new way to automatically determine the number of PCA output components that should be denoised. That is, if a PCA component whose index is $k > k_0$, then this PCA component is selected to perform denoising. We do not denoise the PCA component, otherwise.

The steps to determine k_0 is listed as follows:

Step 1. Let X_b , $b \in [1, B]$, be the PCA output components. We perform the wavelet transform on X_b for one decomposition scale, denoted as Y_b ($b \in [1, B]$), and

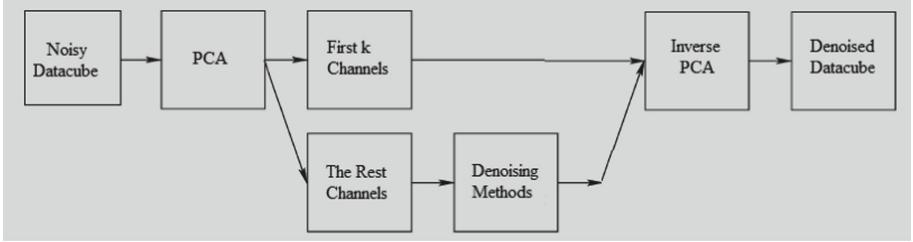


Fig. 1. The flow chart of the proposed denoising technique is depicted here by combining PCA with existing denoising methods. In the figure, the ‘*Denoising Methods*’ can be wavelet shrinkage, BM3D, VBM3D, BM4D, VBM4D, NL-Means, etc.

then estimate the noise standard deviation as [14]:

$$\sigma_n^b = \frac{\text{median}(|Y_{b,1}|)}{0.6745} \quad (2)$$

where $Y_{b,1}$ is the finest wavelet subband.

Step 2. Let $b = 1$ and $T = 0.08$.

Step 3. If $\frac{\sigma_n^b}{\max_k(|X_{b,k}|)} \leq T$, then $b = b + 1$. Go to the beginning of Step 3. Otherwise, go to Step 4.

Step 4. $k_0 = b$. Stop.

We conducted experiments on two simulated hyperspectral imagery data cubes with very low noise level as [5, 6] and [7]. It was concluded that the PCA+BM4D is the best for this level of noise. However, for moderate and high noise levels, PCA+BM4D is not as good as PCA+VBM3D for reducing the noise in hyperspectral imagery. In addition, PCA+VBM3D is much faster than PCA+BM4D in terms of CPU computation time.

3 Experimental Results

We use the same two hyperspectral data cubes as our previous papers [5–7]. The first data cube (Mates et al., 2004) was acquired using the AVIRIS in the Greater Victoria Watershed District (GVWD), Canada, on August 12, 2002. The ground sample distance (GSD) of the data cube was $4 \text{ m} \times 4 \text{ m}$ with nominal AVIRIS SNR of 1000:1. The size of the data cube for testing is 292×121 pixels with 204 spectral bands. Figure 2 shows band #50 of the original noise-free data cube. The second data cube is a simulated hyperspectral data cube that was created from Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data of Cuprite, Nevada, USA. The test noisy data cube was formed by adding normally distributed zero mean noise to achieve a nominal SNR of 600:1. The term ‘nominal SNR’ refers to the ratio of the signal to the noise in the VNIR region in each SNR pattern at certain circumstances [15]. The size of this data cube is 512×614 pixels with 213 spectral bands. Figure 3 shows band #50 of the simulated noise-free Cuprite data cube. In our experiments, we extract a small data cube with 256×256 pixels in each band and 213 bands in total.

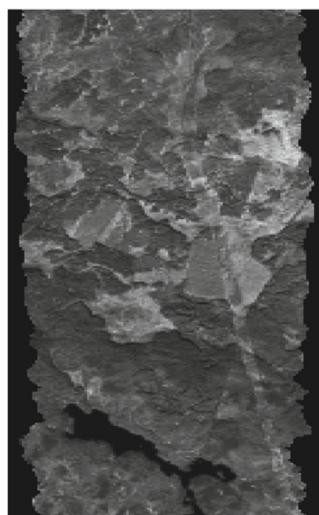


Fig. 2. AVIRIS GVWD scene (spectral band #50).

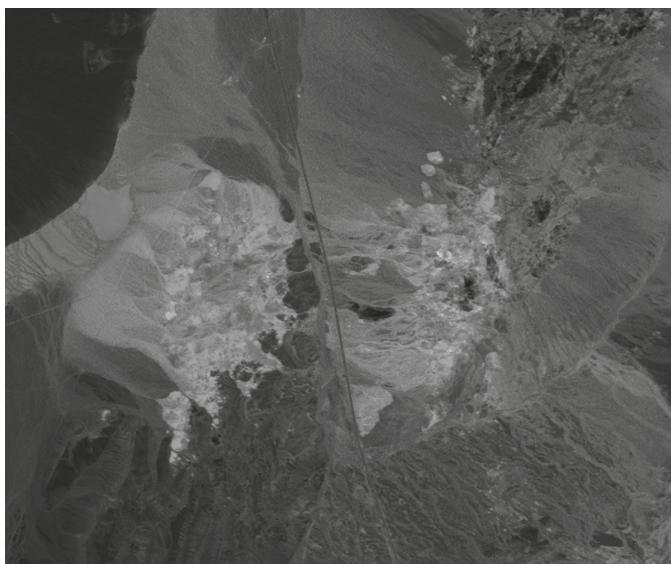


Fig. 3. Simulated Cuprite scene (spectral band #50).

We conducted several experiments by denoising the hyperspectral imagery data cubes directly with and without PCA preprocessing, and we obtained the results as shown in Tables 1, 2 and 3 in terms of signal to noise ratio (SNR). The SNR is defined as

$$SNR(A, B) = \frac{\sum_{i,j,k} A(i, j, k)^2}{\sum_{i,j,k} (B(i, j, k) - A(i, j, k))^2} \quad (3)$$

where i, j, k are the indices of the voxels in the 3D data cubes, B is the test noisy data cube or the denoised data cube, and A is the reference (noise free) data cube. Existing methods without PCA reduce noise less effectively than those with PCA. This is because fine features in the noisy data cubes have been removed as noise. In Table 1, the simulated noisy data cubes have very little noise, which cannot be seen by human eyes. Under such noise condition, the PCA+BM4D is preferable. When the noise levels get higher, the PCA+VBM3D performs the best among all six denoising methods as shown in Tables 2 and 3. In addition, the PCA+VBM3D is much faster than the PCA+BM4D, PCA+VBM4D, and PCA+NL-Means in terms of computation time (See Table 4). It is worthy pointing out that PCA+BM4D and PCA+VBM4D are too slow to be used in practical remote sensing applications. The PCA+Wavelet shrinkage [5] is also a bit slow due to the translation-invariant (TI) signal denoising along the spectral direction. We believe that, by replacing the TI signal denoising with standard signal denoising, the PCA+Wavelet shrinkage should be fast as well.

Table 1. The SNR for the two hyperspectral imagery data cubes with simulated noise.

Datacube	Noisy	Scheme	Wavelet Shrinkage	BM3D	VBM3D	BM4D	VBM4D	NL-MEANS
GVWD	1811.26	No PCA	416.06	541.15	2118.60	3802.52	2817.78	357.56
		PCA	7396.46	9482.32	1927.41	9440.36	9458.43	8435.39
Cuprite	5834.25	No PCA	1938.83	3247.57	7528.58	4072.93	7289.45	1904.84
		PCA	14855.47	14899.68	5967.22	16654.56	16363.59	15017.76

Table 2. The SNR of different denoising methods for the GVWD data cube.

σ_n	Noisy	Scheme	Wavelet Shrinkage	BM3D	VBM3D	BM4D	VBM4D	NL-MEANS
100	249.54	No PCA	379.05	231.47	1908.18	2484.92	1597.74	391.53
		PCA	4643.10	7258.71	7707.11	7132.64	7282.51	6420.89
200	100.11	No PCA	350.68	115.14	1167.85	1430.74	997.37	363.57
		PCA	1856.79	3622.45	3801.09	3660.29	3697.94	2983.36

(continued)

Table 2. (*continued*)

σ_n	Noisy	Scheme	Wavelet Shrinkage	BM3D	VBM3D	BM4D	VBM4D	NL-MEANS
300	67.19	No PCA	369.71	85.88	1026.58	1202.75	853.92	367.52
		PCA	1449.52	2596.05	2737.92	2629.62	2657.91	2116.26
400	53.90	No PCA	404.75	75.62	1008.26	1151.83	813.45	393.35
		PCA	1289.91	2568.99	2738.78	2604.20	2637.00	2007.93
500	46.94	No PCA	445.67	72.76	1036.57	1166.88	815.19	417.66
		PCA	1166.26	2304.15	2477.92	2345.01	2376.75	1805.38
600	42.72	No PCA	488.03	74.11	1089.40	1215.33	839.97	459.76
		PCA	1057.86	2676.10	2907.60	2742.76	2771.97	1945.44
700	39.91	No PCA	531.68	79.00	1156.39	1282.20	879.94	502.33
		PCA	1128.35	2561.56	2824.17	2626.71	2670.88	1840.38
800	37.92	No PCA	573.57	87.62	1230.47	1359.25	930.97	547.00
		PCA	1083.84	2477.75	2743.94	2539.69	2590.79	1780.70
900	36.44	No PCA	612.84	100.12	1308.74	1444.38	990.93	590.26
		PCA	1052.82	2401.89	2690.75	2488.54	2528.88	1727.57
1000	35.29	No PCA	650.04	114.70	1392.40	1533.94	1057.69	631.12
		PCA	1204.32	2325.88	3395.43	1978.20	2185.94	1889.58

Table 3. The SNR of different denoising methods for the Cuprite data cube.

σ_n	Noisy	Scheme	Wavelet Shrinkage	BM3D	VBM3D	BM4D	VBM4D	NL-MEANS
100	257.37	No PCA	1222.32	313.96	3488.12	3680.63	3554.37	1322.00
		PCA	17992.38	17520.79	19384.74	17659.67	18153.02	11565.53
200	105.09	No PCA	1317.20	149.89	2436.18	2459.17	1791.38	1189.42
		PCA	9097.21	8558.03	9687.95	8215.71	8627.10	5719.84
300	71.05	No PCA	1555.80	117.67	2393.16	2408.71	1811.34	1214.89
		PCA	7024.60	6722.91	7784.75	6790.29	7043.23	4045.30
400	57.15	No PCA	1842.48	113.06	2593.12	2575.13	1994.64	1335.77
		PCA	6079.61	5847.30	6856.04	5952.43	6196.96	3341.07

(continued)

Table 3. (*continued*)

σ_n	Noisy	Scheme	Wavelet Shrinkage	BM3D	VBM3D	BM4D	VBM4D	NL-MEANS
500	49.82	No PCA	2150.56	122.75	2878.25	2803.00	2225.80	1429.95
		PCA	5568.88	5321.43	6287.52	5452.69	5676.68	2942.95
600	45.35	No PCA	2465.80	141.16	3194.93	3042.66	2477.39	1516.34
		PCA	5231.04	4964.77	5928.07	5123.63	5342.98	2680.49
700	42.41	No PCA	2782.26	165.24	3524.41	3284.64	2737.23	1597.95
		PCA	5006.38	4733.94	5705.29	4908.35	5115.14	2550.94
800	40.48	No PCA	3104.05	195.00	3870.26	3531.55	3008.10	1702.15
		PCA	4865.63	4569.81	5561.17	4775.25	4968.50	2424.05
900	39.04	No PCA	3414.05	230.03	4217.78	3764.51	3273.62	1750.83
		PCA	4752.94	4457.39	5437.62	4668.18	4857.11	2377.63
1000	37.93	No PCA	3710.77	270.33	4543.46	3982.34	3531.38	1827.01
		PCA	4653.52	4361.55	5355.44	4584.91	4766.16	2301.79

Table 4. The execution time (in seconds) of different denoising methods for the two hyperspectral imagery data cubes.

Data Cubes	Scheme	Wavelet Shrinkage	BM3D	VBM3D	BM4D	VBM4D	NL-MEANS
GVWD	No PCA	1310	1	53	5980	1650	50
	PCA	1350	7.5	58	6000	1660	80
Cuprite	No PCA	2430	3	110	11750	3220	80
	PCA	2520	17.5	125	11950	3390	140

4 Conclusions

Hyperspectral imagery contains hundreds of spectral bands, and it is inevitably contaminated by noise. Reducing noise in a hyperspectral data cube is a challenging task even though there are several methods published in the literature. In this paper, we have studied the effects of PCA on hyperspectral imagery denoising. We combine PCA with wavelet shrinkage, BM3D, VBM3D, BM4D, VBM4D, and NL-Means for hyperspectral imagery denoising. Our experimental results show that PCA+BM4D is good for very

low noise levels and PCA+VBM3D is preferred for moderate and heavy noise scenarios. In addition, VBM3D is much faster than other methods compared in this paper in terms of CPU computational time.

Future research will be conducted for other types of noise, such as signal dependent noise, salt-and-pepper noise, and impulse noise. We would like to investigate the effects of our proposed denoising method on automatic target classification, endmember extraction, and other remote sensing applications. We would also study whether minimum noise fraction (MNF [16]) is better than PCA or not in hyperspectral imagery denoising.

References

1. Ye, M., Qian, Y., Zhou, J.: Multi-task sparse nonnegative matrix factorization for joint spectral-spatial hyperspectral imagery denoising. *IEEE Trans. Geosci. Remote Sens.* **53**(5), 2621–2639 (2015)
2. Yuan, Q., Zhang, L., Shen, H.: Hyperspectral image denoising employing a spectral-spatial adaptive total variation model. *IEEE Trans. Geosci. Remote Sens.* **50**(10), 3660–3677 (2012)
3. Othman, H., Qian, S.E.: Noise reduction of hyperspectral imagery using hybrid spatial-spectral derivative-domain wavelet shrinkage. *IEEE Trans. Geosci. Remote Sens.* **44**(2), 397–408 (2005)
4. Zhao, Y.Q., Yang, J.: Hyperspectral image denoising via sparse representation and low-rank constraint. *IEEE Trans. Geosci. Remote Sens.* **53**(1), 296–308 (2015)
5. Chen, G.Y., Qian, S.E.: Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage. *IEEE Trans. Geosci. Remote Sens.* **49**(3), 973–980 (2011)
6. Chen, G.Y., Qian, S.E., Gleason, S.: Denoising of hyperspectral imagery by combining PCA with block-matching 3D filtering. *Can. J. Remote. Sens.* **37**(6), 590–595 (2011)
7. Chen, G.Y., Bui, T.D., Quach, K.G., Qian, S.E.: Denoising hyperspectral imagery using principal component analysis and block matching 4D filtering. *Can. J. Remote. Sens.* **40**(1), 60–67 (2014)
8. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **16**(8), 2080–2095 (2007)
9. Dabov, K., Foi, A., Egiazarian, K.: Video denoising by sparse 3D transform-domain collaborative filtering. In: Proceedings of the 15th European Signal Processing Conference, EUSIPCO 2007, Poznan, Poland (2007)
10. Maggioni, M., Katkovnik, V., Egiazarian, K., Foi, A.: A nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE Trans. Image Process.* **22**(1), 119–133 (2013)
11. Maggioni, M., Boracchi, G., Foi, A., Egiazarian, K.: Video denoising, deblocking and enhancement through separable 4-D nonlocal spatiotemporal transforms. *IEEE Trans. Image Process.* **21**(9), 3952–3966 (2012)
12. Buades, A.: A non-local algorithm for image denoising. *Comput. Vis. Pattern Recogn.* **2**, 60–65 (2005)
13. Jolliffe, T.: Principal Component Analysis. Springer-Verlag, New York (2002). <https://doi.org/10.1007/b98835>
14. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**(3), 425–455 (1994)
15. Mates, D. M., Zwick, H., Jolly, G., Schulten, D.: System studies of a small satellite hyperspectral mission, data acceptability, Macdonald, Dettwiller, Assoc., Richmond, BC, Canada, Can. Gov. Contract Rep. HYTN-51–4972 (2004)

16. Green, A.A., Berman, M., Switzer, P., Craig, M.D.: A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Trans. Geosci. Remote Sens.* **26**(1), 65–74 (1988)



Two-Channel VAE-GAN Based Image-To-Video Translation

Shengli Wang¹, Mulin Xieshi², Zhangpeng Zhou¹, Xiang Zhang², Xujie Liu¹,
Zeyi Tang², Yuxing Dai³, Xuexin Xu³, and Pingyuan Lin³(✉)

¹ Maintenance Company of State Grid Power Company in Gansu Province, Lanzhou 730000, Gansu, China

² State Grid Info-Telecom Great Power Science and Technology Co., LTD., Fuzhou 350000, China

³ School of Informatics, Xiamen University, Xiamen 361005, China

linpy@stu.xmu.edu.cn

Abstract. We propose a VAE-GAN network with a two-channel decoder for addressing multiple image-to-video translation tasks, i.e., generating multiple videos of different categories by a single model. We consider this image-to-video translation as a video generation task rather than a video prediction that needs multiple frames as input. After training, the model only requires the first frame of the video and its corresponding attribute to generate the required video. The advantage of combining the Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) is to avoid the shortcomings of both: VAE components can give rise to blur, and unstable gradients caused by the GAN. Extensive qualitative and quantitative experiments are conducted on the MUG [1] dataset. We draw the following conclusions from this empirical study: compared with state-of-the-art approaches, our approach (VAE-GAN) exhibits significant improvements in generative capability.

Keywords: Video generation · Variational autoencoder · Generative adversarial network

1 Introduction

Image-to-video translation has become an important and widely studied research area in computer vision. It has resulted in many exciting applications in multimedia content generation. By inputting only one image and video target attribute, such as the emotion when we generate facial expression video, the translation system can generate the video related to the input. We consider this translation task to be part of the video generation task, harder than the video prediction task which requires multiple inputs.

Recently, several studies [3, 12, 16, 25, 27, 29] have used Variational Autoencoder (VAE) structures to model motion stochasticity in video for generating different possible future frames. The VAE uses pixel-level loss functions such as the Mean Square Error (MSE) to minimize the log-likelihood and maximize the variational upper bound to make the training tractable. However, the pixel-level loss will cause the model to predict

results that just correspond to the average, i.e., it will generate blurry samples. Besides the VAE, Generative Adversarial Networks (GAN) [7] have emerged as a promising framework in video generation [6, 15, 18, 20, 23, 24]. By distinguishing blurry samples from natural ones via adversarial learning, the GAN can obtain realistic synthetic results compared with the VAE. However, as theoretically investigated in [2], the GAN still suffers from unstable gradient and mode collapse, which limits its application in video generation with complex actions.

As mentioned above, the GAN can achieve more realistic results than the VAE, while the VAE is more robust to model collapse than the GAN. This has led to attempts to combine the merits of the GAN and the VAE to achieve higher levels of image quality while avoiding unstable gradients. For example, [13] and [30] use the VAE-GAN structure to implement both the text-to-video and text-to-image tasks. More recently, [11] simply adds a discriminator of GAN onto the top of a VAE, leading to interesting results in the video prediction task. However, in practice, this method leads to generating images that are still not of high quality.

In the work described in this paper, we further explore the potential of combining VAE and GAN for multiple image-to-video translation tasks, aiming to synthesize multiple high-quality videos by only providing the first frame of the video sequence together with the target video category.

We claim that it is hard for the image generator in existing VAE-GAN methods [11, 13] to generate videos of high quality. Video-generation is an intrinsically highly difficult task since our image-to-video task has only one input, unlike most video prediction tasks which require multiple inputs to provide motion information. In order to solve this problem, we propose a dual-channel VAE decoder that improves the modeling capabilities of the image generator in the future. Specifically, we introduce an auxiliary decoder channel (fine optimization), it is with the content information extraction module (Rough optimization) connection, the video is generated in a rough to fine. Therefore, the VAE decoder assembly generates a varied but blurred video on a rough level, and then connects the auxiliary decoder channel and the content extraction channel to refine the blurred results to obtain fine details.

As a result, the VAE component produces diverse but blurry videos at a coarse level, and then the GAN component refines the blurry results to obtain fine details.

In summary, our contributions are three-fold:

1. We propose a novel two-channel decoder framework, to perform multiple image-to-video translation tasks by only providing both the first frame of the video and the target category.
2. We design two loss functions (the identity feature matching loss and the connected feature matching loss) for our image-to-video model to enhance the quality of generated videos and stabilize the training of the overall framework.
3. Both qualitative and quantitative results demonstrate the superiority of our proposed model when compared with the state-of-the-art alternatives.

2 Proposed Approach

In this section, we introduce the VAE component and the loss when we train the VAE component in our VAE-GAN structures. We then further design and describe two loss functions for our training model. With these loss functions at hand, the VAE-GAN model can generate videos that preserve both the semantic contents and the overall structure of the object. We compare VAE-GAN architecture with VAE or GAN architecture for the image-to-video translation task in Fig. 1. For generating the video sequence $V = \{I_0, I_1, \dots, I_T\}$, where T is the total number of frames in the video, we use two essential components, namely a) video generation, b) generator loss function.

Specifically, the process of blurry video generation firstly utilizes 2D convolution layer E_c to extract semantic content from the first frame I_0 of the input video sequence V . The semantic content is denoted by tensors of different sizes f_c . Then, we sample the motion information z from the Gaussian models which mean and variance are outputted by the VAE encoder E_m when training the VAE component. We utilize 3D convolution layers to constitute the VAE encoder E_m . When we train the GAN component, we do not input the video V into encoder E_m to get motion information z but sample it from the prior Gaussian models $z \sim p_\theta(z)$, which means is zero and variance is one.

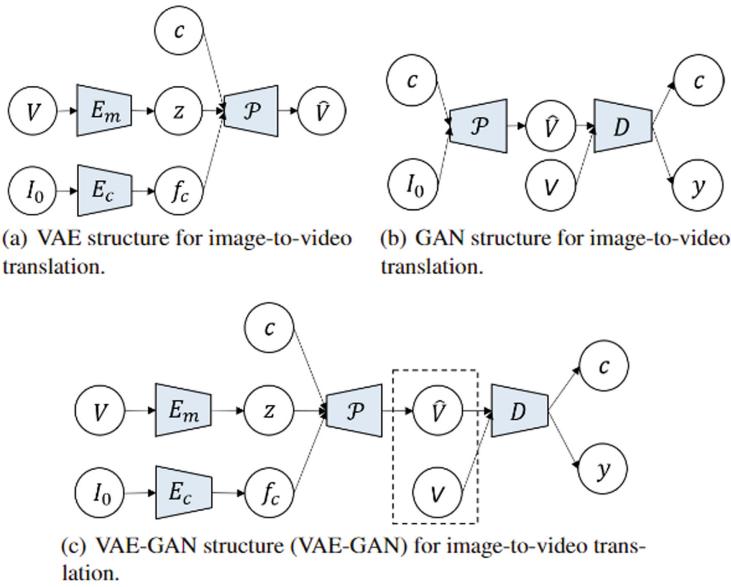


Fig. 1. Illustration of the structure of the VAE [3, 29], GAN [6, 20], VAE-GAN network (VAE-GAN) for image-to-video translation task. V represents the input video, I_0 means the first frame of the input video V . z is the latent variable sampling from the VAE encoder E_c when training VAE component or from the prior when training GAN component, c is the attribute of target generated video, and y is a binary output, representing whether the video is real or fake. In the model of VAE, GAN, and the ‘VAE-GAN’, the output video is denoted as \hat{V} .

In Fig. 1, c is denoted as the attribute of target generated video, e.g., the emotion when we generate facial expression video. In the experiment, we represent c as a one-hot vector, for example $c = [100000]$, where the first digit 1 indicates that the target generation facial expression video should display a specific expression. By changing the position of the 1, the model can generate six different expression videos. Finally, we utilize the motion information z , semantic content feature f_c , and a one-hot vector c as the input to the decoder network P to reconstruct the video \hat{V} in the VAE-GAN. The decoder network P is illustrated in Fig. 2.

The unit marked D in Fig. 1 denotes the discriminator in our VAE-GAN structure and contains two components. The first is the set of 2D convolution layers D_i used to distinguish a single frame between generated frame and real frame, and the second is the set of 3D convolution layers D_v used to distinguish videos between generated video and real video. The output of D is a binary variable y , which represents whether the video is real or fake.

The GAN generator loss used in the blurry video generation is introduced in Sect. 2.2. The generator loss function stabilizes the gradient convergence, maintains the consistent high-level features between the generated and real videos, and plays a role in preserving the overall structure of the object by building a connection between the VAE and GAN components.

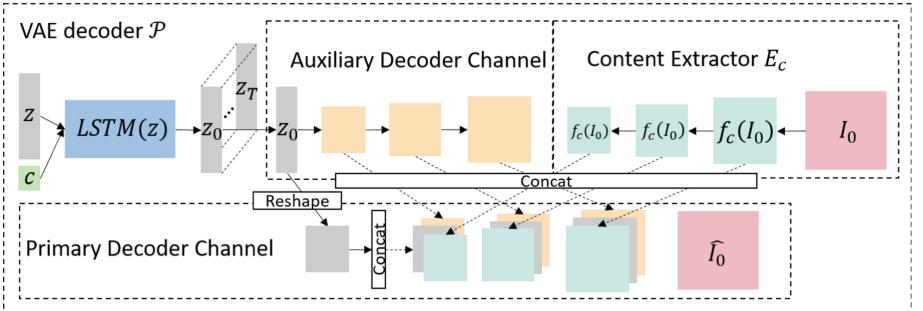


Fig. 2. Illustration of our two-channel VAE decoder structure. Here, we take how to generate the first frame of the video \hat{I}_0 as an example.

2.1 Video Generation

To generate video, following [23], we assume that the video generation task can be decomposed into two sub-processes, namely a) motion generation and b) content information generation. Thus, this section of the paper focuses on modeling motion and content features.

VAE Loss. Firstly, following the VAE approach, we aim to train a variational neural network $q_\theta(z|V)$ to approximate the otherwise intractable latent posterior $p_\theta(z|V)$. We sample z from the Gaussian models parameterized by the mean and variance outputted from the VAE encoder $E_m(V)$. For the sake of simplicity, this process is re-formulated

as $z = E_m(V)$ in this work. E_m is a 3D convolutional network. Assuming that the prior $p_\theta(z)$ is known (e.g., Gaussian), we utilize a neural network to approximate a complex likelihood $p_\theta(V|z)$, which maximizes the data likelihood $p_\theta(V)$. In our image-to-video task, for reconstructing the video, given the first frame I_0 and the video attribute c , our proposed decoder is formulated as:

$$\hat{V} = P(z, f_c, (I_0), c) \quad (1)$$

where the vector z is obtained by $z = E_m(V)$ in the VAE training process, $f_c(I_0) = E_c(I_0)$, and \hat{V} is the reconstructed blurred video. The 2D convolution network E_c is used to extract the semantic information contents from the first frame I_0 . The corresponding content features $f_c(I_0)$ are the output of each convolutional layer of the network E_c . The semantic content thus consists of several tensors of different sizes. When implementing the decoder, the content features are connected to the decoder P in the manner of the skip-connect [17].

The decoder P consists of an LSTM [2] network and several convolutional and deconvolutional layers, and we will detailed describe the decoder P next subsection. In conclusion, when we train the VAE component in our VAE-GAN structures, the network aims to minimize the following loss:

$$L_{VAE} = D_{KL}(q_\phi(z|V)||p_\theta(z) + \|\hat{V} - V\|_1 \quad (2)$$

where D_{KL} is the Kullback-Leibler (K-L) divergence. Using Eq. (2). we can minimize the KL-divergence between the posterior $q_\phi(z|V)$ and the prior $p_\theta(z)$, and minimize the L_1 reconstruction loss between the generated video \hat{V} and the real video V , i.e., $\|\hat{V} - V\|_1$.

Two-Channel VAE Decoder Structure P . To constrain our model to utilize the motion information contained in the sampled latent variable z , we have designed a new two-channel VAE decoder structure. The first channel is the auxiliary decoder channel, which outputs different deconvolution features according to z_0 . The second channel is the primary decoder channel. In each layer the features consist of the previous output of the deconvolution, the content features come from E_c and the features from the auxiliary decoder channel. The initial input for the primary decoder channel is obtained by reshaping z_t . The two-channel VAE decoder P is illustrated in Fig. 2; for each sampled latent variable z , either coming from the VAE encoder $z = E_m(V)$ when we train the VAE component or sampling from the prior $z \sim p_\theta(z)$ when training GAN component, we first feed z and its video attribute c into the LSTM network [21] to yield $z_0 \sim z_t$, where t represents the number of video frames. We also obtain the content features $f_c(I_0)$ from the content extractor E_c , $f_c(I_0) = E_c(I_0)$. Finally, the two-channel VAE decoder can convert the latent variable $z_0 \sim z_t$ and content features $f_c(I_0)$ to the video with t frames, which is denoted as $\hat{V} = \hat{I}_0 \sim \hat{I}_t$.

Perceptual Loss. Drawing on the recent use of the perceptual loss for super resolution [9], here the perceptual loss is employed to suppress feature difference between the output of VAE decoder P blurred video \hat{V} and the real video V . It is defined as follows:

$$L_{Perceptual} = \sum_i \|\psi_i(\hat{V}) - \psi_i(V)\| \quad (3)$$

For training VAE component, we input the real video V to VAE decoder E_m for sampling motion vector z , input first video frame I_0 to content extractor E_c for content feature f_c . Then input the z, f_c and video attribute c into the VAE decoder P to reconstruct blur video \hat{V} . The loss we use is VAE loss Eq. 2 and perceptual loss Eq. 3. For training GAN component, we sample motion vector z from the prior $z \sim p_\theta(z)$, and combine it with content features f_c and video attribute c into the VAE decoder P to reconstruct blur video \hat{V} as well.

2.2 Generator Loss Functions

For enhancing the spatiotemporal consistency of the video and avoiding mode collapse, we first revisit the feature matching loss for generating images [19] in VAE-GAN. We design two feature matching loss functions, respectively termed as a) the identity feature matching loss and b) the connected feature matching loss. These both have the effect of stabilizing the gradient.

Identity Feature Matching Loss. Although the feature matching loss L_{FM} to some extent stabilizes the gradient, the feature matching loss cannot play a perfect role in retaining the high-dimensional features of the video and its overall spatial structure. Furthermore, considering that our task is to generate video by feeding the VAE-GAN an image, we re-formulate the loss L_{FM} to give a new loss L_{IFM} as follows:

$$L_{IFM} = \sum_n \left(\frac{1}{2} \sum_t \left\| \psi_{-1}(I_t^n) - \psi_{-1}(\hat{I}_t^n) \right\|_1 + L_{FM}(V^n, I_0^n, z) \right) \quad (4)$$

By minimizing the distance between the high-level features, L_{IFM} can improve the consistency of both the generated video and real video content information. This also prevents the mode collapse. We utilize the identity feature matching loss only when we train the GAN component, and the gradient will affect the content extractor E_c , VAE decoder P in the VAE-GAN.

Connected Feature Matching Loss. A connected feature matching loss is utilized to increase the generated quality of the video. Specifically, the connected feature matching loss builds a connection between the VAE and GAN components, forcing the model to approximate the high-level features between the input video and generated video. As a result, the connected feature matching loss plays a role in preserving the overall structure of the object. In the VAE training process, motion vector z is obtained by $z = E_m(V)$, so z depends on the video V uniquely. Therefore, L_{CFM} is further simplified to:

$$\begin{aligned} L_{CFM} &= \frac{1}{2} \sum_t \left\| \psi_{-1}(I_t) - \psi_{-1}(\hat{I}_t) \right\|_2^2 + \frac{1}{2} \|f_{DC}(V) - f_{DC}(G(E_m(V), I_0, c))\|_2^2 \\ &\quad + \frac{1}{2} \sum_t \left\| f_{DI}(I_t) - f_{DI}(\hat{I}_t) \right\|_2^2 \end{aligned} \quad (5)$$

I_0 is the first frame of real video V , and I_t and \hat{I}_t are the t -th frames of real video V and generated video ($G(E_m(V), I_0, c)$), respectively. Here, Euclidean distance is used to

calculate the loss in the VAE. The blur problem will be addressed since we minimize the distance between the high-level features rather than at the pixel level. It should be noted that the gradient generated by the loss L_{CFM} controls the updates in the entire VAE and GAN framework.

2.3 Final Objective

The loss function for our VAE-GAN is defined as the weighted sum of the individual losses described above. The final objective function is summarized as follows.

$$L_{Full} = L_{DC} + \lambda_1 L_{VAE} + \lambda_2 L_{Perceptual} + \lambda_3 L_{IFM} + \lambda_4 L_{CFM} \quad (6)$$

where $\lambda_1 = 30$, $\lambda_2 = 5$, $\lambda_3 = 5$ and $\lambda_4 = 5$, they are the hyper-parameters which depend on each other.

To provide a bridge between the VAE and GAN structures, the pixel-level loss which is the L_1 reconstruction loss in Eq. 2 and GAN loss Eq. 4, 5 are used in combination. These losses will influence the shared component: content extractor E_c and VAE decoder P . Here, the pixel-level loss prevents the occurrence of mode collapse by constructing different videos from the different latent variables. The GAN loss, on the other hand, to some extent addresses the recurring blur problem.

3 Experiments

In this section, we briefly introduce the datasets and the evaluation metric used in our experiments described in Sect. 3.1. Then we introduce the details of our implementation and the baselines used for comparison in Sect. 3.2. Next, our qualitative and quantitative results are presented and compared to three state-of-the-art approaches in Sect. 3.3. These are used to validate the effectiveness of our proposed method.

3.1 Datasets and Evaluation Metrics

MUG Dataset MUG [1] contains 86 subjects. For the experiments, we utilize videos containing more than 67 frames. Each video has one of six different facial emotion categories: anger, disgust, fear, happiness, sadness, and surprise. We form our dataset by sampling a sequence of sub-frames from each video. According to the video duration, we sample different numbers of sub-frames, but the number of sub-frames does not exceed four. The length of each sub-frame sequence is 16 frames, and each sequence contains facial expressions that transition to the most prominent part of the expression, and then gradually decrease in expression level.

Following [23], we cropped the face regions and scaled them to 96 * 96 pixels to form our facial expression dataset. The total number of training and test dataset is 1487 and 285, respectively.

Evaluation metrics Average content distance: For quantitative comparison, the Average Content Distance (ACD) indicator is used to measure the content consistency of the generated facial expression videos [23]. In order to extensively evaluate the proposed

methods, three extensions of ACD [20] are used: (1) ACD-I; (2) ACD-C; (3) ACD-G. ACD-I calculates the average feature distance between the generated frame and the input frame; (2) ACD-C calculates the feature distance between all possible frame pairs in one video; (3) ACD-G calculates the average frame-to-frame feature distance between the generated frame and the real frame.

Intrinsically, the smaller the value of ACD, the more consistent the video content and the stronger the model generation capability of the network architecture. Following [23], we use a pre-trained Open Face neural network [4] to extract the feature vector of the human face.

PSNR and SSIM: Like [3, 27], we use the PSNR and SSIM to evaluate the proposed methods when generating the videos.

Fréchet Inception Distance: Fréchet Inception Distance (FID) [8] is used to calculate the similarity between the generated image or video sequence and the target image or video sequence. The lower the FID value, the better the generated model is. Following [26], we used ResNeXt [28] and I3D [5] as an inception network to extract the features for frames and video sequences, respectively. For the sake of simplicity, we term them as FID-ResNeXt and FID-I3D, which focus on the visual quality of the images and video sequences, respectively. Additionally, we used the pre-trained OpenFace neural network [4] as the inception network to calculate FID. We refer to it as FID-OpenFace, which mainly focuses on judging the visual quality of the facial images.

3.2 Experimental Setup

In this section, we introduce the implementation of the network architecture and the details of the procedure used for network training.

Network Architecture: For the motion encoder E_m , our design involves five Conv3D layers, one Conv2d layer, and one linear operator to extract the latent variable z with 512 hidden units. For the content extractor E_c , our design has four Conv2d layers to extract the input image features, and the channel for each extracted feature has 32, 64, 128, and 256 units, respectively. Each hidden unit of the output of the LSTM $z_0 \sim z_t$ in the decoder P has 1024 units. We reshape each channel to size 4×4 with 64 channels as the initial input for the primary decoder channel. The primary decoder channel consists of four 2D deconvolutional modules and two 3D convolution layers. The auxiliary decoder channel has four 2D deconvolutional modules. Each 2D deconvolutional module consists of two 2D convolution layers and an up-sampling operator. For D_v , we used five Conv3D layers, and four Conv2d layers for D_i .

Training Details: We used the Adam optimizer [10] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In our experiments, the total number of trained batches was set as 150,000 for the MUG dataset. The batch size was set as 2 for this dataset. This means that the total number of epochs is about 202 for MUG. In our experiments, the learning rate is fixed until the 100,000 batches for the MUG dataset and reduced by a factor of 100 at each epoch. For the more stable gradient obtained from the discriminator, the least-squares loss [14] is implemented for L_{DC} .

Baseline Algorithms: In our experiments, we compare our algorithm both qualitatively and quantitatively with a number of state-of-the-art alternatives including MoCoGAN [23], AffineGAN [20], and P2PVG [27]. The first two are based on the GAN, while the third is based on VAE. Their architectures and the corresponding experimental setup are described as follows:

MoCoGAN [23] can generate videos, unconditionally, or conditionally. In this experiment, we used the conditional image-to-video mode, given the first video frame as input. We used their recommended parameters for training using the MUG and Weizmann datasets. In total 12,000 batches are used with batch size 16, which means that the total number of epochs is about 130 for MUG, 68 for Weizmann Action-I, and 57 for Weizmann Action-II. For the Weizmann dataset, the number of human action categories is 10.

AffineGAN [20] is a video-generated model that uses an action variable representation to control the expression intensity of each generated frame. We train the model for 600 epochs using the reported default parameters.

P2PVG [27] controls the video generation process using the start and end frames. Thus, we regard the result of P2PVG as the achievable ceiling of video generation performance with a single input image. In our experiment, we aim to observe the visual quality gap between our architectures and P2PVG. The number of epochs is 200, the epoch size is 200, and the batch size is 32. The remaining parameters are as reported in [27].

3.3 Qualitative and Quantitative Results

Qualitative Results. Figure 3 compares our generation results with three baseline approaches when generating facial videos according to the target emotion category:

- For the AffineGAN model, we train a specific model for each emotion category. As seen in Fig. 3, it is seen that AffineGAN failed to reveal the significant changes in muscle motion, e.g., “Disgust” (in the row (g)). It is explained that (1) AffineGAN is caused by the unstable training of the GAN network for obtaining the optimal generator; (2) AffineGAN is trained for each frame, instead of the whole video, which causes the AffineGAN model to neglect the continuity of the video.
- Similar to AffineGAN, for P2PVG, six models are trained for six different emotions. As seen in Fig. 3, the P2PVG based on the VAE model generates many identical frames without significant muscle motion change. For example, the “Happy” video sequence in Fig. 3 (in the row (d)) shows P2PVG generated the same frames under $T = 3$, $T = 5$, and $T = 7$ or under $T = 11$, $T = 13$, and $T = 15$. Some of the generated start and end frames are entirely consistent with the input, indicating that the model moves the training target to the consistency of the first and last frames without paying attention to the consistency of the video.
- MoCoGAN can synthesize continuous facial expression videos. However, it is seen that the quality of some frames of the video is poor, and some differences between the identity of the generated video and the input frame have existed. For example, the

row (c) and the row (a) ‘‘Happy’’ video sequence in Fig. 3 show a difference in human identity. The results indicate that MoCoGAN only needs to consider generating real action videos without considering the identity of the input frame and the generated video.

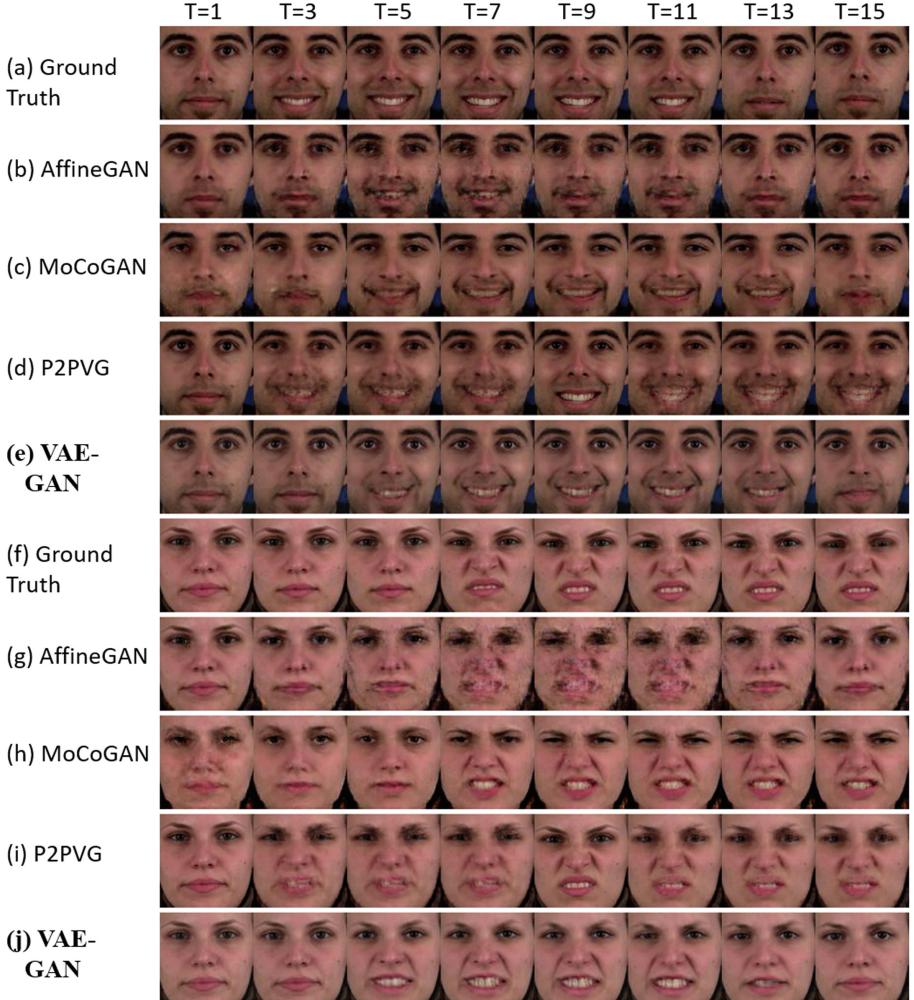


Fig. 3. Qualitative comparison results on the MUG dataset. The video sequences in the row (a)–(e) and row (f)–(j) show the happy and disgusted expressions, respectively

Compared with the state-of-the-art algorithms, VAE-GAN produces more coherent sequences of facial movements when generating facial expression videos. The results demonstrate that our designed loss and structure can stabilize the gradient and suppress the identity difference between the fake and the real videos (Fig. 4).

Table 1. Comparison of image-to-video generation approaches on the MUG dataset. The best result is in bold. The blue means the second-best result.

(a) Subtable 1					
Model	FID-ResNeXt	FID-I3D	FID-OpenFace		
MoCoGAN	0.47	80.18	0.129		
AffineGAN	0.79	95.25	0.076		
P2PVG	0.43	80.13	0.043		
VAE-GAN	0.21	74.91	0.048		

(b) Subtable 2					
Model	PSNR	SSIM	ACD-I	ACD-C	ACD-G
MoCoGAN	22.309	0.578	0.509	0.374	0.608
AffineGAN	22.444	0.614	0.505	0.423	0.583
P2PVG	25.339	0.741	0.524	0.393	0.442
VAE-GAN	24.310	0.690	0.401	0.334	0.468

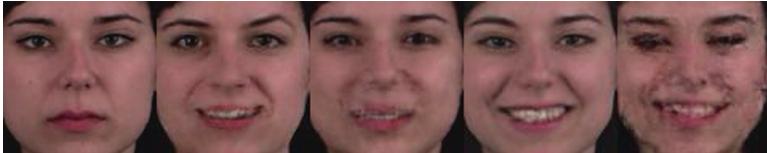


Fig. 4. A screenshot of the evaluation video for human performance. The far left of the evaluation video is the input frame. To the right of the input frames are four videos generated by different models according to the input frame.

Quantitative Results. Table 1 reports the results of applying FID, PSNR, SSIM, and ACD to the MUG dataset. The VAE-GAN architectures that use both our newly designed loss and our two-channel network structure outperform the alternative state-of-the-art approaches in terms of FID-ResNeXt, FID-I3D, ACD-I, and ACD-C. However, they are still inferior to the P2PVG model in terms of FID-OpenFace, PSNR, SSIM, and ACD-G. The poor performance when compared with P2PVG in terms of the FID-OpenFace, PSNR, SSIM, and ACD-G metrics is due in part to the experimental setup used. To simplify the generation of video using the P2PVG model, we input two sets of start and end frames to the model. The first set is utilized to generate the facial emotion expression video and these transition from indistinct expressions to prominent expressions. We then input a second set to generate a video transitioning from a prominent expression to an indistinct expression. Finally, we concatenate the two generated frame sequences to form the complete facial emotion video. We therefore use four frames to generate a complete emotion video. These frames significantly improve the final performance metrics. Another reason for the larger PSNR of the P2PVG model is that the loss function of P2PVG includes MSE, which directly optimizes the value of PSNR and makes the calculated value of PSNR larger. Additionally, as observed in Fig. 3, the P2PVG model

generates incoherent motion videos, and the visual appearance is poor. Therefore, ‘VAE-GAN’ still has a competitive performance edge over the P2PVG model on the facial expression generation task.

4 Conclusions

In this paper, we have investigated the challenging problem of how to generate videos by using a starting frame and a target class label. We propose a two-channel VAE-GAN structure for generating high-quality and coherent videos. Furthermore, we design a novel identity feature matching loss and a connected feature matching loss for our VAE-GAN network. The former stabilizes the gradient and makes the generated video frames approach the real-world video frames in terms of their high-level feature contents. Moreover, the latter loss plays an essential role in preserving the overall structure of the object in the generated frames. To demonstrate the superiority of our two-channel VAE-GAN, we perform extensive experiments on the MUG facial emotion dataset. According to our comparative results, the VAE-GAN model outperforms the competitive state-of-the-art methods on the image-to-video translation task.

References

1. Aifanti, N., Papachristou, C., Delopoulos, A.: The mug facial expression database. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, pp. 1–4 (2010)
2. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: 5th International Conference on Learning Representations, ICLR, Toulon, France, 24–26 April 2017 (2017)
3. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
4. Baltrušaitis, T., Robinson, P., Morency, L.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Compute Vision, WACV, Lake Placid, NY, USA, 7–10 March 2016, pp. 1–10 (2016)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 21–26 July 2017, pp. 4724–4733 (2017)
6. Fan, L., Huang, W., Gan, C., Huang, J., Gong, B.: Controllable image-to-video translation: a case study on facial expression generation. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, pp. 3510–3517 (2019)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680. Curran Associates, Inc. (2014)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30, pp. 6626–6637 (2017)
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. Computer Vision – ECCV 2016, pp. 694–711 (2016). https://doi.org/10.1007/978-3-319-46475-6_43

10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015 (2015)
11. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. CoRR (2018)
12. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.: Flow-grounded spatial-temporal video prediction from still images. In: Computer Vision - ECCV 2018 - 15th European Conference, pp. 609–625 (2018). https://doi.org/10.1007/978-3-030-01240-3_37
13. Li, Y., Min, M.R., Shen, D., Carlson, D.E., Carin, L.: Video generation from text. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), pp. 7065–7072 (2018)
14. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 2813–2821 (2017)
15. Nam, S., Ma, C., Chai, M., Brendel, W., Xu, N., Kim, S.J.: End-to-end time-lapse video synthesis from a single outdoor image. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, June 16–20, 2019, pp. 1409–1418 (2019)
16. Pan, J., et al.: Video generation from single semantic label map. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 3733–3742 (2019)
17. Ronneberger, O., P.Fischer, Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
18. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: IEEE International Conference on Computer Vision ICCV Venice, Italy, 22–29 October 2017, pp. 2849–2858 (2017)
19. Salimans, T., et al.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, vol. 29, pp. 2234–2242 (2016)
20. Shen, G., et al.: Facial image-to-video translation by a hidden affine transformation. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2505–2513 (2019)
21. 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. In: Advances in Neural Information Processing Systems, vol. 28, pp. 802–810 (2015)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, pp. 1–14. Computational and Biological Learning Society (2015)
23. Tulyakov, S., Liu, M., Yang, X., Kautz, J.: Mocogan: decomposing motion and content for video generation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, 18–22 June 2018, pp. 1526–1535 (2018)
24. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances in Neural Information Processing Systems, vol. 29, pp. 613–621. Curran Associates, Inc. (2016)
25. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: Computer Vision – ECCV 2016 - 14th European Conference, pp. 835–851 (2016). https://doi.org/10.1007/978-3-319-46478-7_51
26. Wang, T.C., et al.: Video- to-video synthesis. In: Advances in Neural Information Processing Systems, vol. 31, pp. 1144–1156. Curran Associates, Inc. (2018)
27. Wang, T., Cheng, Y., Lin, C.H., Chen, H., Sun, M.: Point-to-point video generation. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, Seoul, Korea (South), 27 October–2 November 2019, pp. 10490–10499 (2019)
28. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 21–26 July 2017, pp. 5987–5995 (2017)

29. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: Advances in Neural Information Processing Systems, vol. 29, pp. 91–99. Curran Associates, Inc. (2016)
30. Zhang, C., Peng, Y.: Stacking VAE and GAN for context-aware text-to-image generation. In: Fourth IEEE International Conference on Multimedia Big Data, BigMM, Xi'an, China, 13–16 September 2018, pp. 1–5 (2018)



High-Voltage Tower Nut Detection and Positioning System Based on Binocular Vision

Zhiyu Cheng¹, YiHua Luo¹, JinFeng Zhang¹, Zhiwen Gong², Lei Sun^{2(✉)}, and Lang Xu³

¹ State Grid Anhui Electric Power Co., Ltd., Hefei 230022, Anhui, China

² Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
pchen@ahu.edu.cn

³ National Engineering Research Center for Agro-Ecological Big Data Analysis and Application, School of Internet, Anhui University, Hefei 230601, China

Abstract. Ultra High Voltage (UHV) transmission is the most advanced transmission technology in the world. However, it is difficult for the daily maintenance of high voltage power towers. Based on the development of robots and in-depth learning, this paper proposes a visual-based pylon climbing robot to detect high-voltage tower nuts. An improved yolov5 is developed by adding coordinate attention (CA) module to the backbone, and assigning different weights to different levels of features, replacing the Concat of neck species with Full-Concat. Experiment results showed that our proposed scheme can detect and locate nuts very well, and our trained model can also be well applied in our devices.

Keywords: Binocular vision · YOLOv5 · Coordinate attention mechanism · Coordinate attention

1 Introduction

UHV transmission is the most advanced transmission technology in the world. High voltage transmission is an important component of the power system because of its large capacity, low loss, long distance and other advantages. High-voltage power transmission equipment is basically outdoors, which may lead to equipment damage and nut loosening due to natural environment, material aging, rainwater and other factors. Therefore, regular maintenance and inspection of the high-voltage tower is required to eliminate hidden dangers. Traditional methods are generally manual inspection, but there are some drawbacks in manual inspection, such as high risk, low efficiency, high labor intensity and high rate of missed inspection.

Therefore, based on the safety, efficiency, cost, and other considerations of high voltage power tower patrol inspection, as well as the gradual popularity of industrial robots, some researchers proposed to use robots instead of manual patrol inspection. Robot patrol refers to the fact that the robot can climb on the power tower independently, and with the help of vision, the hidden dangers of the power tower can be eliminated, and

real-time data can be returned to the hands of ground workers. Compared with traditional manual patrol inspection, robot patrol inspection has the advantages of high efficiency, low miss rate and good safety.

In order for the patrol robot to complete its work, it must accurately detect the position information of the nuts and obstacles on the climbing route, so it needs to use vision to complete its work. Currently, the mainstream are monocular vision and binocular vision. Monocular vision uses a single camera to locate the target [1]. The target depth information is derived mainly from the two-dimensional characteristics of the image. Binocular vision [2] is similar to human visual system, has higher measurement speed and accuracy, and has the advantages of simple structure and easy use, so it is widely used.

At present, deep learning algorithms have good results in the field of computer vision target detection. Especially after the YOLO [3] series algorithm was proposed, because of its fast speed and high accuracy, deep learning algorithms are gradually applied to the industrial environment. This paper combines the YOLOv5 algorithm and binocular vision algorithm to detect the nut, and calculates the position information of the nut based on the detection.

2 Principle of Nut Detection

Nut recognition and positioning are based on nut detection. In this paper, the improved YOLOv5 algorithm is used to realize the real-time detection of nuts. The nut detection process in this article is shown in Fig. 1.

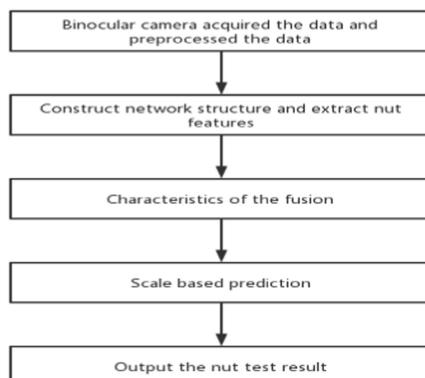


Fig. 1. Nut detection process

2.1 YOLOv5 Model Detection and Recognition

YOLOv5 is improved on the basis of YOLOv3, and has higher detection speed and accuracy compared with YOLOv3. YOLO5 includes four different network structures,

YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The depth and width of these four network structures are different. In order to meet the speed requirements, this paper chooses the YOLOv5s network model with the smallest network depth and width [4]. The whole network structure is composed of input, backbone, neck and detection, with the focus on backbone and neck. As shown in Fig. 2.

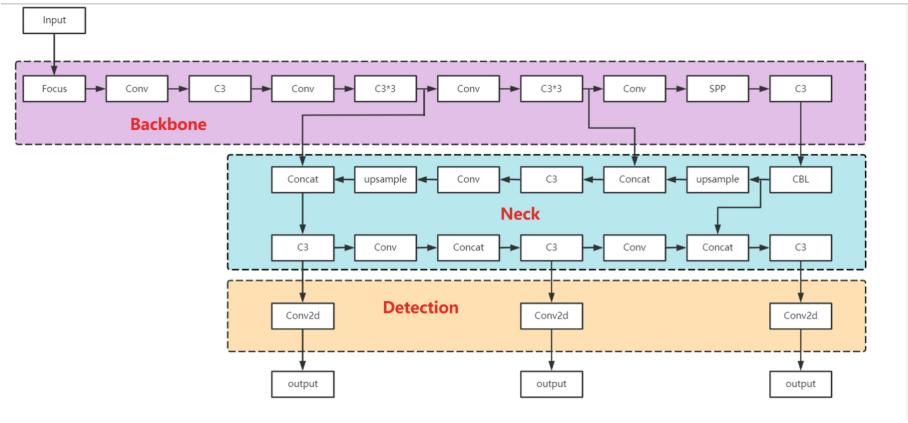


Fig. 2. YOLOv5 network structure

The input part of YOLOv5 mainly includes the following three parts: first, data enhancement based on Mosaic algorithm. Four images are divided into a group, and one image is synthesized by random scaling, splicing, or superposition for training, which can enhance the number of small targets in the data set and thus improve the model's detection ability of small target objects. The second is the adaptive anchor frame calculation. In the model training, YOLOv5 will start the adaptive anchor frame calculation function according to the parameters, and adaptively calculate the best anchor frame value in different types of training sets. The difference between the output prediction frame and the real frame will be compared to calculate the difference, and then to reverse update the iterative network parameters. The third is adaptive picture scaling, the original picture to a unified standard size, and then sent to the network for detection.

Backbone of the model is mainly composed of Focus, C3 and Spatial pyramid pooling (SPP) [5] modules. Focus is a slicing operation for feature map, which divides the data into 4 pieces, and each piece of data is equivalent to two times of sampling. The longitudinal channel is spliced, and then the convolution operation is carried out. The specific process is shown in Fig. 3. The C3 module is similar to the ResNet structure [6], and the main idea is to skip connections. The overall operation does not change the size of the input, but the model gains stronger learning performance. In the feature generation part, YOLOv5 adopts the SPP network of YOLOv3 to complete it, and integrates multi-scale receptive fields through down-sampling of different sizes to improve feature extraction ability.

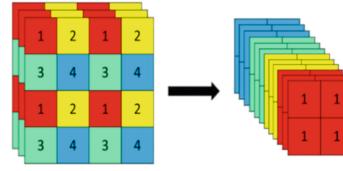


Fig. 3. Focus slice operation

Based on Feature Pyramid Networks (FPN) [7] structure and PANet [8] network, Neck module realizes multi-scale Feature fusion network of FPN + PAN structure. It can enhance the network's ability of feature fusion for objects of different scale. FPN mainly improves the detection results of small target objects by integrating high- and low-level features. Based on FPN, PAN adds a bottom-up information flow path, shortening the information transmission path, and aims to enhance accurate low-level location information to the whole network.

2.2 The Proposed Method

As the nut target in the nut image is small, occupies few pixels, and is easily affected by the background, the original model algorithm is easy to lose the characteristic information of the small target during the convolution sampling, and the detection effect is not good for the small target. An improved YOLOv5 algorithm is proposed to introduce the coordinate attention mechanism into the feature extraction module of the YOLOv5 network. It enables the backbone network to focus on the region of interest, and improves the ability of feature description. For small targets and dense targets, the feature information can be effectively extracted to further improve the accuracy of detection. In addition, we propose to replace the Concat module of the feature fusion layer with the Full-Concat module. This model learns important features from different inputs, focuses on important features and neglects less important features. The network structure is shown in Fig. 4.

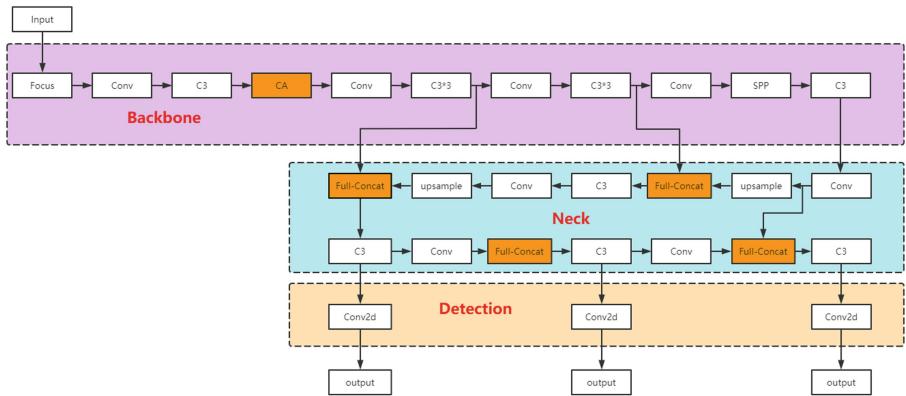


Fig. 4. The network structure of improved YOLOv5.

1) The CA Net

Currently, commonly used attention mechanisms include SE and CBAM, etc. But to some extent, they all focus on channel information and ignore location information. HOU *et al.* [9] proposed coordinated attention (CA), which effectively introduced channel information into the attention mechanism. The feature expression ability of the convolutional neural network model is effectively improved. The specific structure is shown in Fig. 5.

The CA module can be divided into two parts: coordinate information embedding and coordinate information attention generation. Specifically, two one-dimensional global pooling operations aggregate input features along with vertical and horizontal directions into two separate feature graphs for a specific direction. Then, two feature maps containing specific direction information are generated and two attention maps are generated. Each attention map can obtain the feature information range of the target extending in a specific direction in the input feature map. Then the two feature maps are multiplied onto the input feature map to enhance the expressive power of the feature map. The resulting feature map contains the location information of the target.

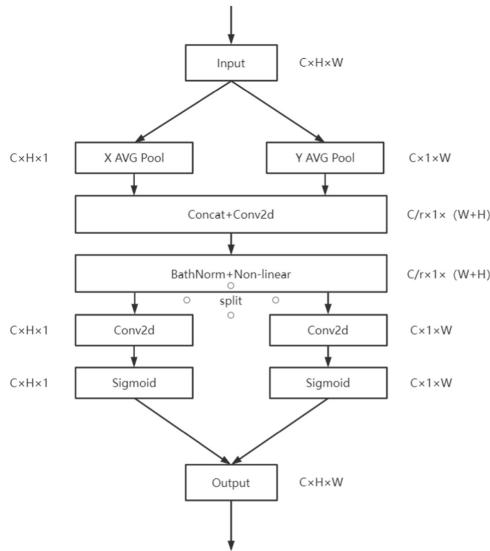


Fig. 5. Coordinate attention.

2) The Full-Concat Block

YOLOv5 network uses Concat module for feature fusion. All levels of functionality in this module are treated equally, but the different levels of functionality are linked together. However, different levels of input characteristics affect output characteristics. Therefore, in this article we consider assigning different weights to features

of different levels. Referring to the work of Google [10], this article replaced the Concat module in YOLOv5 with the full Concat module. In Full-Concat, it uses the fast normalized fusion method. The formula is shown in (1).

$$O = \sum_i \frac{\omega_i}{\epsilon + \sum_j \omega_j} \cdot I_i, \quad (1)$$

where ω_i represents the learnable weight, O indicates the output feature graph, I_i represents feature maps of different levels of inputs and ϵ is a small value to avoid instability.

2.3 Principle of Nut Positioning

The positioning of the binocular stereo vision is based on the principle of parallax. Two pictures of the same object are taken from different positions by two cameras, and the coordinate values of the object in a three-dimensional space are obtained by calculating the parallax of the nuts in the two pictures based on triangular geometry.

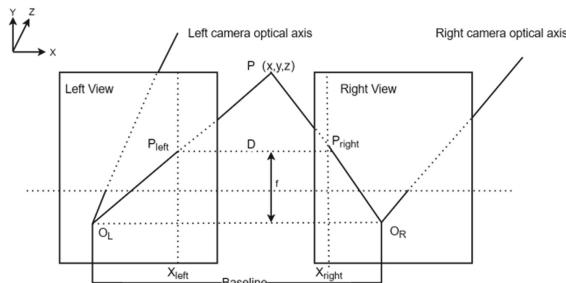


Fig. 6. Binocular stereo imaging schema

Figure 6 is a schematic diagram of binocular parallax. The center of the two cameras is O_L , O_R , and the target point $P(x, y, z)$ is viewed at the same time. Baseline is the connecting distance between the projection centers of the two cameras and the camera focal length is f . The image coordinates of their images in the left and right cameras are $P_{left} = (x_L, y_L)$, $P_{right} = (x_R, y_R)$. Now, assuming that the images from both cameras are on the same plane, the Y coordinates of P_{left} and P_{right} are the same, that is, $y_L = y_R = Y$. From the triangular geometric relationship.

$$x_L = \frac{fx}{z}, x_R = \frac{f(x - B)}{z}, Y = \frac{fy}{z}. \quad (2)$$

The distance between the coordinates of the left and right camera images is called the parallax D , and the parallax has $D = x_L - x_R$. From this, it can be calculated that the three-dimensional coordinates of the target P in the camera coordinate system are.

$$x = \frac{Bx_L}{D}, y = \frac{BY}{D}, z = \frac{fB}{D}. \quad (3)$$

Since the camera focal length f and baseline B are internal parameters in the camera, the three-dimensional coordinates of any point on the left camera can be determined in the camera coordinate system, as long as the right camera finds its corresponding matching point.

2.4 Evaluating Indicator

The main evaluation indicators for the YOLO model are: Precision, recall, mAP_0.5 and mAP_0.5:0.95. The precision is defined as (4), The recall is defined as (5).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (5)$$

True Positions (TP) refers to the correct positive samples assigned; True Negatives (TN) refers to the correctly allocated negative samples; False Positions (FP) refers to the positive samples wrongly allocated; False Negatives (FN) refers to the negative samples incorrectly allocated. Average precision refers to the curve composed of recall as the horizontal axis and precision as the vertical axis. The area enclosed by the curve is AP and its formula is as follows:

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i)p(r_{i+1}), \quad (6)$$

The mean average precision (mAP) refers to the mean value formula of the average accuracy of all categories in the dataset, as shown in the following figure:

$$mAP = \frac{1}{m} \sum_{i=1}^m AP_i, \quad (7)$$

where M represents the number of samples in the test set.

3 Experiment and Result

3.1 Experimental Environment

The operating system of this experiment is Ubuntu 16.04, The CPU model is Intel(R) Xenon(R) with refresh rate 2.4 GHz, 6 cores, logical processor numbers are 12 and the running memory is 32G. The GPU model is Geforce GTX 1080ti, the size of the video memory is 11G, and the size of the memory is 32G. The deep learning framework is Python 1.8.0 and the experimental language is Python 3.8.0, CUDA version is 10.2.

3.2 Datasets

The data in this paper are collected manually in the outdoor electric tower by means of a binocular camera. When collecting data, we fully considered the influencing factors such as illumination and background, collected a total of 1800 nut data, manually marked them through LabelImg, and saved the pictures in YOLO's VOC format. The data includes a single nut and dense nuts at the connecting plate of the tower, and we added some screw hole pictures to interfere with the experiment. We randomly divided the data set, into training set, test set and verification set according to 7:2:1 (Fig. 7).

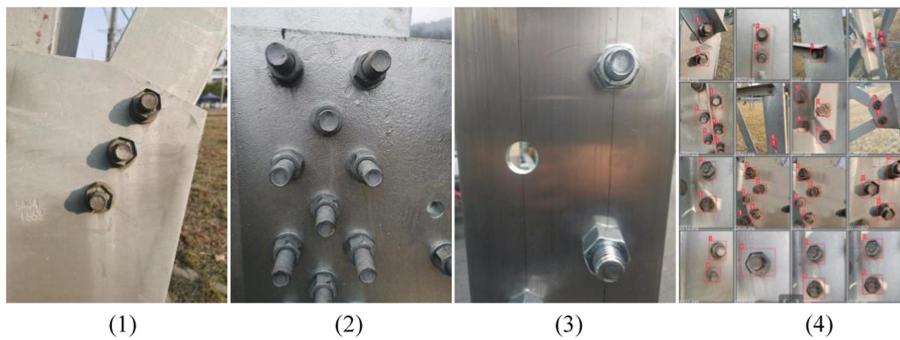


Fig. 7. The figure shows: (1) single nut; (2) nuts on the connecting plate; (3) Add the data of the screw hole; (4) Some renderings of training

3.3 Results and Analysis

For better detection results, this paper deals with epochs and batch_size did a comparative test. Batch_size is set to 16, and based on this epochs are set to 100, 200, 300 respectively. The result is that 95.6% mAP can be obtained when epochs is set to 300. When batch_size is set of 32, epochs to 100, 200, 300, and 91.3% of the mAP can be obtained when the epochs are 300. The experimental results are shown in Fig. 8. Some parameters of the training part of this experiment are set: batch_size is 16, the initial learning rate is 0.01, epoch is set to 300, and the pre_training model is YOLOv5s pt, and CA module are added after the first C3 module in the backbone. The nut in the recognition image can be obtained more accurately. As shown in the Fig. 9, the detection results of the original yolov5 algorithm and our algorithm are shown.



Fig. 8. Experimental results with different batch_size and epochs

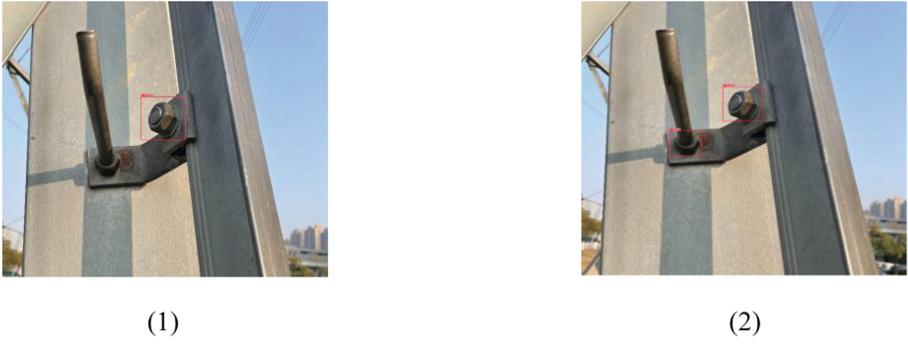


Fig. 9. The figure shows: (1) Yolov5 algorithm detection results (2) improved yolov5 algorithm detection results

To fully demonstrate the superiority of this algorithm, we have compared the original yolov5 algorithm with the yolov3 algorithm using the same data set and parameter settings. The experimental results are compared as shown in Table 1.

Table 1. Comparison of results with other networks.

	Precision(%)	Recall(%)	mAP_0.5(%)
YOLOv3	91.6	90.4	89.7
YOLOv5	93.8	93.4	94.8
OURS	94.9	92.8	95.6

As shown in Table 1, the mAP of the improved algorithm is up to 95.68%, which is better than the original YOLOv5 algorithm and higher than the YOLOv3 algorithm. Our improved algorithm achieves 94.97% more precision than the original YOLOv5 and YOLOv3, but our algorithm is slightly inferior to the original yolov5 algorithm in recall. The experimental results show that our proposed algorithm has a good effect on small targets such as nuts and can accurately identify nuts in complex scenes.

To verify the reliability of the model, the algorithm is used to locate the center point of the nut. Our experiment was carried out with a Kinect camera and a binocular camera, where Kinect camera is used mainly to roughly locate the approximate position of the center point of the nut, and then binocular camera is used to precisely locate the nut, so that the position information of the nut can be accurately obtained. The structure of the experimental equipment is shown in Fig. 10. The nut is fixed in the experiment and mainly depends on the moving binocular camera. The initial position of the camera is set to (0,0,200 mm) and the distance between the binocular camera and the nut is not changed during the experiment. We move only 10 mm in the X, Y direction each time. The experimental results are shown in Table 2. By comparing the experimental results recorded for each movement with the actual results, we can see that our positioning error is between 1 and 3 mm, so our experimental model has certain reliability and accuracy.

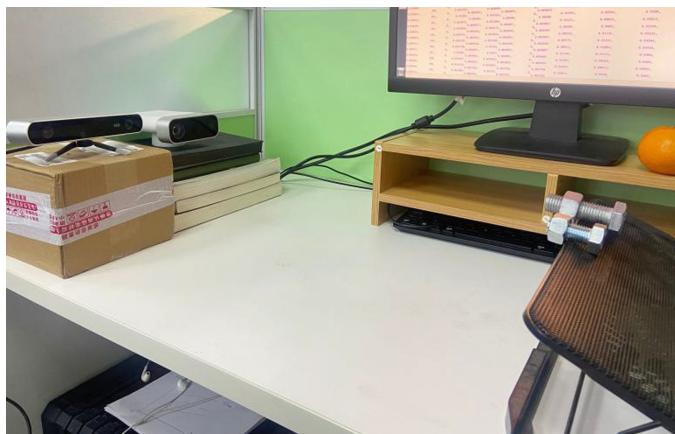


Fig. 10. The structure of the experimental equipment

As shown in Fig. 11, “1” represents the label of the nuts and “0.82” represents the confidence, x and y are the coordinates relative to the binocular camera, and dis is the distance from the binocular camera.



Fig. 11. Nut position information obtained by the Zed camera

Table 2. The experimental results of nut positioning (the data unit is mm)

Sequence Number	Nut real Position (x, y, z)	Nut measurement position (x, y, z)	Error (x, y, z)
1	(0, 0, 200)	(0, 0, 200)	(0, 0, 0)
2	(10, 0, 200)	(10, 0, 200)	(0, 0, 0)
3	(20, 0, 200)	(19, 8, 199)	(0, 2, 1)
4	(30, 0, 200)	(28, 1, 200)	(2, 1, 0)

(continued)

Table 2. (continued)

Sequence Number	Nut real Position (x, y, z)	Nut measurement position (x, y, z)	Error (x, y, z)
5	(40, 0, 200)	(40, 0, 199)	(0, 0, 1)
6	(50, 0, 200)	(47, 1, 200)	(3, 1, 0)
7	(50, 10, 200)	(50, 10, 200)	(0, 0, 0)
8	(50, 20, 200)	(49, 19, 200)	(1, 1, 0)
9	(50, 30, 200)	(50, 29, 200)	(0, 1, 0)
10	(50, 40, 200)	(50, 38, 199)	(0, 2, 1)
11	(60, 50, 200)	(60, 50, 199)	(0, 0, 1)
12	(70, 60, 200)	(69, 59, 199)	(1, 1, 1)
13	(80, 70, 200)	(80, 69, 199)	(0, 1, 1)
14	(90, 80, 200)	(88, 80, 199)	(2, 0, 1)
15	(100, 90, 200)	(98, 90, 199)	(2, 0, 1)

4 Conclusion

To meet the requirements of climbing robots for better inspection of high voltage towers, In this article, based on the yolov5 algorithm, a Coordinate Attention (CA) module is added to the backbone for nut recognition. And we will put the trained model in the edge computing device JETSON and use the ZED binocular camera to detect and locate the nut. Our device works well in the outdoor environment and achieves the desired results, providing good visual assistance to the climbing robot device.

Acknowledgement. This work was supported by the State Grid Anhui Electric Power Co., Ltd. (No. 5212002000AS).

References

1. Zixing, C.: Fundamentals of Robotics. China Machine Press, Beijing (2009)
2. Haihua, F.U.: Identification of handwriting by writing robot. Sci. Technol. Winds **19**, 4 (2019)
3. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: Unified, realtime object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788 (2016)
4. Yang, X., Jiang, W., Yuan, H.: traffic sign recognition and detection based on Yolov5. Inf. Technol. Inform. (4), 28–30 (2021)
5. Ouyang, W., Zeng, X., Wang, X., et al.: DeepID-net: deformable deep convolutional neural networks for object detection. IEEE Trans. Patt. Anal. Mach. Intell. **39**, 1 (2016)
6. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708 (2017)

7. Lin, T.Y., Dollar, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 936–944 (2017)
8. Liu, S., Qi, L., Qin, H., et al.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
9. Hou, Q.B., Zhou, D.Q., Feng, J.S.: Coordinate attention for efficient mobile network design. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
10. Haibo, L., Shanli, T., Shuang, S., Haoran, L.: An improved yolov3 algorithm for pulmonary nodule detection. In: 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), vol. 4, pp. 1068–1072 (2021)



Palmprint Recognition Using the Combined Method of BEMD and WCB-NNSC

Li Shang¹(✉), Yuze Zhang², and Zhan-li Sun³

¹ Department of Communication Technology, College of Electronic Information Engineering, Suzhou Vocational University, Suzhou 215104, Jiangsu, China
s10930@jssvc.edu.cn

² Department of Civil, Environmental and Geomatic Engineering, Institute of Finance and Technology, University College London, London, England
zctpyz3@ucl.ac.uk

³ Department of Automation, College of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, China

Abstract. A novel image palmprint reconstruction method using the combined method of bi-dimensional empirical mode decomposition (BEMD) and weight coding based non-negative sparse coding (WCB-NNSC) is proposed here. The BEMD algorithm is especially adaptive for non-linear and non-stationary 2D-data analysis. And the weight coding based NNSC algorithm includes more class information than that of the basic NNSC. For each original palmprint image, its first two order high frequency intrinsic mode functions (IMFs) extracted by BEMD are denoised by Wiener filter, then denoised IMFs and low frequency IMFs are fused by weighted method and normalized, moreover, using these preprocessed images as test images of WCB-NNSC, and feature basis vectors can be successfully learned. Moreover, using suitable classifiers, the palmprint recognition task can be implemented. Further, in the same experimental condition, compared with palmprint feature recognition methods of standard ICA and NNSC, simulation results show that our method proposed in this paper is indeed efficient and effective in performing palmprint recognition task.

Keywords: BEMD method · Weight coding · Non-negative sparse coding (NNSC) · Palmprint images · Feature recognition

1 Introduction

At present, palmprint recognition technique offers a promising future for medium-security access control system. And some of methods rooted in neural computation, such as Principal Component Analysis (PCA) [1], Independent Component Analysis (ICA) [1], Non-negative Matrix Factorization (NMF) [2], and Non-negative Sparse Coding (NNSC) [3, 4], etc., have been used widely image feature extraction and recognition task [5–7]. Above-mentioned algorithms are all the data-adaptive representations, which are tailored to the statistics of data. However, in feature extraction task, PCA is not suitable to process high dimension data, and ICA, NMF and NNSC all ignore the important

factor of classification. So, in order to implement well classification task, in this paper, considered class information, we proposed the Weight Coding based NNSC (WCB-NNSC) algorithm. At the same time, in order to save the high and low frequency information, the Bi-dimensional Empirical Mode Decomposition (BEMD) [8–12] is used to preprocess palmprint images. For each original palmprint image, the intrinsic mode functions (IMFs) extracted by BEMD are fused by weighted method and normalized. Then these preprocessed images are used as test images, at the same time, considering the image patch segmentation method, the training set of the WCB-NNSC algorithm can be obtained. Furthermore, palmprint features can be extracted successfully by using the combined transform method. Then, using classifiers selected, the palmprint recognition task can be implemented efficiently. Moreover, compared with the basic NNSC algorithm, simulation results also show that our palmprint recognition method discussed in this paper is efficient in fact application.

2 The BEMD Algorithm

The BEMD algorithm was proposed on the basis of the EMD approach, which is a signal analysis method behaving the discriminating capacity of high time-frequency [8] and is adaptive and appears to be a suitable for nonlinear, non-stationary data analysis. But the EMD method is fit to process 1D signal. In order to apply EMD on images, the BEMD method was developed [8–10] to be used in image processing field. The BEMD method extracts zero-mean 2D AM-FM components called Intrinsic Mode Functions (IMF). Each IMF has the same number of zero crossings and extrema. Given an image $I(x, y)$, the decomposed result of BEMD is written as:

$$I(x, y) = \sum_{i=1}^K IMF_i(x, y) + Ir(x, y) \quad (1)$$

where IMF_i denotes the i th level IMF component, Ir is the residual term, and the initial image is set as $I(x, y)$ with the size of $M \times N$ pixels. Compute the 2D upper and lower envelope by connecting extrema points with radial basis functions, and then the maximum and minimal envelope extrema, denoted by S_{upper} and S_{lower} respectively, are obtained. And the mean envelope E_m can be obtained. namely $E_m = (S_{upper} + S_{lower}) / 2$. Then, utilized the precision of $Ires_j(x, y) = Ir - E_m$ to train the end condition. If $Ires_j$ doesn't meet the accuracy requirement, let $Ir = Ires_j$, otherwise, let $Ir_j = Ir_{j-1} - Ires_j$. The standard deviation (SD) rule is usually defined as follows [8]:

$$SD = \sum_{x=0}^M \sum_{y=0}^N \frac{\|Ires_j(x, y) - Ir(x, y)\|^2}{Ir^2(x, y)} < \alpha \quad (2)$$

In the decomposition process of BEMD, the number of decomposed levels is not too large. Here, the residual level is not considered, and we design an adaptive weighted value rule to extract information of detail levels, which is defined as follows:

$$Ire = w IMF_1 + w(1 - w) IMF_2 + \cdots + w(1 - w)^{L-2} IMF_{L-1} \quad (3)$$

where the weighted coefficient $w = (L - 2)/L - 1$, and the fusion image can be obtained:

$$\tilde{I} = [Ire - \text{Min}(Ire)] / [\text{Max}(Ire) - \text{Min}(Ire)] \quad (4)$$

Thus, the BEMD image can be obtained by using Eq. (4). The fusion multimodal components behave good complementarity to describe the global and detail characteristics of the original image.

3 Weight Coding Based NNSC Algorithm

3.1 The Cost Function

The basic NNSC algorithm doesn't care about the existence of different classes, and lack the property of being class-specific. To avoid above defects, referring to the document [3], the WCB-NNSC algorithm is defined as the following form:

$$E_w = \frac{1}{2} \sum_i \left\| \mathbf{x}_i - \sum_m w_m s_{mi} \right\|_2^2 + \gamma \sum_{i,m} f(s_{mi}) + \beta \sum_m \sum_{\substack{i, \tilde{i} \\ q(i) \neq q(\tilde{i})}} \frac{\mathbf{w}_m^T \mathbf{x}_i \mathbf{w}_m^T \mathbf{x}_{\tilde{i}}}{n_{q(i)} n_{q(\tilde{i})}} \quad (5)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iN})^T$ ($i = 1, 2, \dots, L$) denotes the samples, $\mathbf{x}_{\tilde{i}}$ denotes another labeled sample column vector, $s_{mi} = (s_{1i}, s_{2i}, \dots, s_{Mi})$ ($m = 1, 2, \dots, M$) denotes the sparse coefficients, and the weights was denoted by $\mathbf{w}_m = (w_{m1}, w_{m2}, \dots, w_{mN})^T$. $q(i)$ and $q(\tilde{i})$ are respectively the label of \mathbf{x}_i and $\mathbf{x}_{\tilde{i}}$. $n_{q(i)}$ and $n_{q(\tilde{i})}$ are the number of samples corresponding to the class of \mathbf{x}_i and $\mathbf{x}_{\tilde{i}}$. The weights $\mathbf{x}_i, \mathbf{w}_m$ and row vector s_m subject to the non-negative constraints.

In Eq. (3), the sparse punitive function $f(s_{mi})$ is defined the negative logarithm of coefficients' sparse density, namely $f(s_{mi}) = -\log(p(s_{mi}))$. The density $p(\cdot)$ is selected as follows [6]:

$$p(s) = \frac{1}{2b} \frac{(d+2)[0.5d(d+1)]^{(0.5d+1)}}{\left[\sqrt{0.5d(d+1)} + |s/b| \right]^{(d+3)}} \quad (6)$$

where parameters $d, b > 0$, d is a sparsity parameter and b is a scale parameter. Parameters d and b are estimated by Eq. (5).

$$\begin{cases} b = \sqrt{E\{s^2\}} \\ d = \frac{2-k+\sqrt{k(k+4)}}{2k-1} \\ k = b^2 p_s(0)^2 \end{cases} \quad (7)$$

The sparser the cost function is, the more the activation is spread over different s_{mi} . The influence of the sparsity term is scaled wit the positive constant γ . The last term in Eq. (3) is weight coding term of different class information, which causes cost if a weight coding \mathbf{w}_m has a large inner product with differently labeled samples \mathbf{x}_i and $\mathbf{x}_{\tilde{i}}$. The influence of the weight term is scaled with the positive constant β .

3.2 Training Rules

The minimization of the cost functions of sparse coefficient and weight coding can be done in turn by applying coefficient and weight steps. The updating rule of coefficient vectors s_m is implemented by using an asynchronous fixed-point search, at the same time, w_m vectors are kept constant. In the same way, when updating weight vectors w_m , sparse coefficient vectors s_m are kept constant. The updating formula of s_m and w_m are defined as follows:

$$\begin{cases} s_m = \lambda \left(w_m^T x_i - \sum_{\tilde{m} \neq m} s_{\tilde{m}i} w_{\tilde{m}}^T w_m \right) \left(w_m^T w_m \right)^{-1} - \lambda \gamma f'(s_{mi}) \left(w_m^T w_m \right)^{-1} \\ w_m = \lambda w_m - \lambda \eta \left(\sum_i \sum_{\tilde{m}} s_{\tilde{m}i} w_{\tilde{m}} s_{mi} - \sum_i x_i s_{mi} \right) + \lambda \eta \beta \sum_{\substack{i, \tilde{i} \\ q(i) \neq q(\tilde{i})}} \frac{x_i (w_m^T x_{\tilde{i}})}{n_{q(i)} n_{q(\tilde{i})}} \end{cases} \quad (8)$$

where $\lambda > 0$ and $w_m = / \|w_m\|_2$. This update rule is applied to randomly chosen s_{mi} until convergence. The weight step is a single gradient step with a fixed step size η .

4 Experimental Results

4.1 Palmprint Feature Extraction

In test, 600 palmprint images from 100 persons were selected from the Hong Kong Polytechnic University (PolyU) database. Each person has 6 images coming from respectively

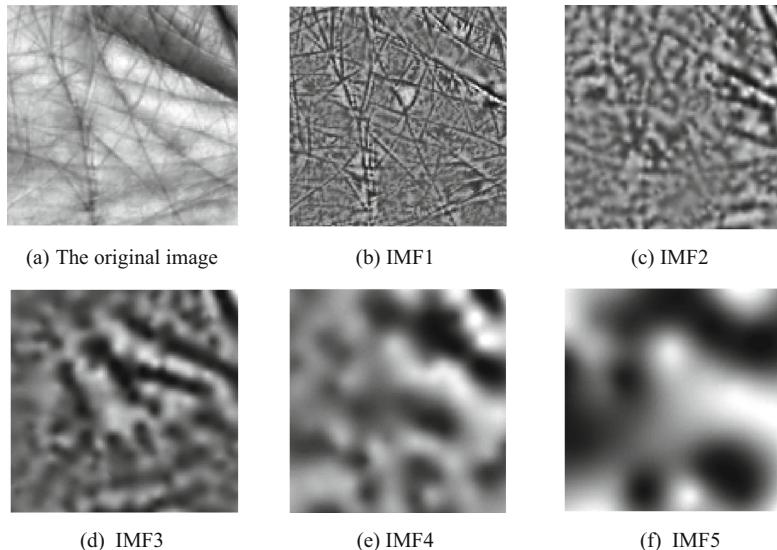


Fig. 1. The original image and its five IMFs obtained by BEMD algorithm.

the left palm and the right palm. Each palmprint image is decomposed into five IMF components from high frequency to low frequency by BEMD, as shown in Fig. 1. The image preprocessed by BEMD is shown in Fig. 2.

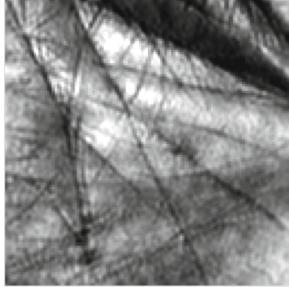


Fig. 2. The preprocessed image by the BEMD algorithm.

The First two order high frequency IMFs are firstly denoised by Winner filter and the denoised results. Then for denoised IMFs and low frequency IMFs, the weighted fusion of IMFs and the preprocessed image are respectively implemented according to Eq. (3) and Eq. (4). To reduce the computational cost, each preprocessed image is scaled to 64×64 pixels. Then, in 600 preprocessed images, each person's first three images are selected as training samples, other three images are treated as testing samples, and each image is converted a column vector. Thus, the training set X_{train} and test set X_{test} are obtained which have the same size of 300×4096 pixels. And in the set of X_{train} , each image is randomly sampled Q times with an $p \times p$ pixel patch, and the input set \hat{X}_{train} of WCB-NNSC with the size of $p^2 \times 300Q$ can be obtained. Then considering different feature dimensions, the WCB-NNSC feature bases can be trained. Here, considering the length limitation of this paper, the WCB-NNSC feature basis images with 64 dimension are given, which is shown in Fig. 3.

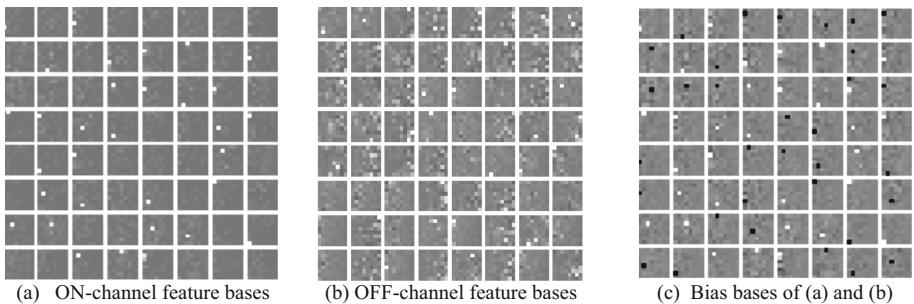


Fig. 3. Basis vectors obtained by the algorithm combined BEMD and WCB-NNSC.

As a contrast, in the same way, let the input set \hat{X}_{train} be the input set of WCB-NNSC, then the NNSC feature bases can be learned, which were shown in Fig. 4. In Fig. 3 and

Fig. 4, the white represents the positive pixels, the gray denotes the zero pixels, and the black shows the negative pixels. Clearly, feature vectors obtained by the combined method of BEMD and WCB-NNSC have more distinct sparseness and locality than those obtained by WCB-NNSC.

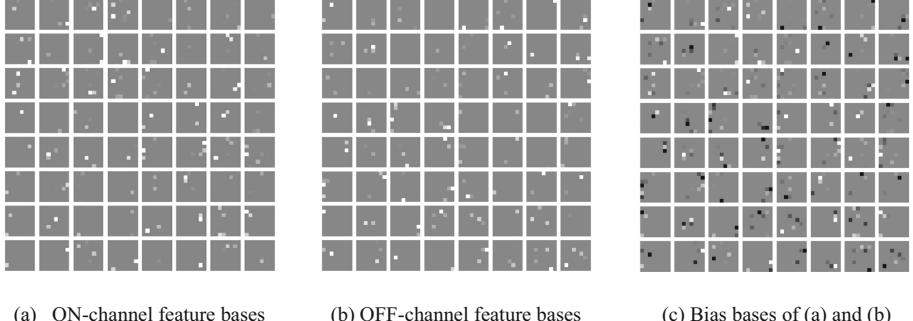


Fig. 4. Basis vectors obtained by WCB-NNSC.

4.2 Feature Recognition

For palmprint images preprocessed by BEMD, referring to the framework IIproposed in the document [1], the palmprint recognition task can be implemented. This method's goal is find the statistically independent feature coefficients. The pixels are variables and input images are observation. First, the optimal number of principal components (PCs) can be obtained by using PCA and three different classifiers, such as the simple Euclidean distance, the probabilistic neural network (PNN) and the radial basis probabilistic neural network (RBPNN) [12]. The recognition results of the three types of classifiers are shown Table 1. In Table, it is clear to see that when the classifier is fixed, as the number of PCs increases, the recognition rate does not increases. When PCs' number exceeds certain value, the recognition either decreases or remains unchanged. Considering the computation speed and the recognition precision, the suitable number of PCs is selected as 90. Namely, in WCB-NNSC and NNSC training, the input matrix's size is selected as 90×4096 pixels. Using the same of three classifiers, the recognition rates of our method proposed in this paper are shown in Table 2. At the same time, the compared results obtained by standard NNSC and WCB-NNSC are also listed in Table 2. From Table 2, when the recognition method is fixed, it is easy to see that, the recognition result obtained by RBPNN classifier is the highest in three classifiers. By contrary, when the classifier is fixed, the recognition rate of our method proposed here is the best. This also proves that our recognition method and RBPNN classier are indeed efficient in palmprint recognition.

Table 1. Recognition rate of PCA using three types of classifiers.

PCs	Euclidean Distance (%)	PNN (%)	RBPNN (%)
49	82.00	83.67	93.33
64	83.67	85.33	95.33
75	84.33	86.33	95.70
80	84.67	86.00	96.20
90	85.00	86.67	98.13
110	85.00	86.33	98.13
115	85.00	86.00	98.13

Table 2. Recognition rate obtained by different algorithm and classifiers (PCs = 90).

Algorithm	Euclidean distance (%)	PNN (%)	RBPNN (%)
NNSC	90.15	92.92	93.85
WCB-NNSC	92.82	93.75	95.53
BEMD-WCB-NNSC	94.36	96.56	98.87

5 Conclusion

A palmprint recognition method combined BEMD and WCB-NNSC is discussed in this paper. The information of high frequency and low frequency features of palmprint images are obtained by using the BEMD method. Further, considered the denoising of the first two high frequency IMFs and the weighted fusion of all IMFs, the pre-processed image set can be obtained, which contain more details of palmprint images. Further, palmprint features were extracted successfully by WCB-NNSC algorithm, at the same time, considering different classifiers, the recognition task was implemented easily. Compared with other methods, the simulation results show that the our algorithm is the best corresponding to the same classifier. Similarly, corresponding to each algorithm, the recognition property of RBPNN is the best than other classifiers used here. In a word, the experimental results prove our palmprint recognition method proposed is indeed efficient in application.

Acknowledgement. This work was supported by the grants of National Science Foundation of China (No. 61972002).

References

1. Connie, T., Teoh, A., Goh, M., Ngo, D.: Palmprint recognition with PCA and ICA. *Image Vis. Comput.* **NZ** *3*, 227–232 (2003)
2. Hoyer, P.: Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **5**, 1427–1469 (2004)
3. Shang, L., Zhou, Y., Sun, Z.: Image recognition using local features based NNSC model. In: The 13th International Conference on Intelligent Computing, pp. 190–199, Liverpool, UK (2017)
4. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996)
5. Zhang, X.: Bingzi LYU: Gabor-2DPCA palmprint recognition based on improved BEMD. *J. Xi' Univ. Posts Telecommun.* **23**(2), 40–48 (2018)
6. Liu, Z., Peng, S.: Boundary processing of bidimensional EMD using texture synthesis. *IEEE Signal Process. Lett.* **12**(1), 33–36 (2005)
7. Ding, S., Du, P., Zhao, X., Zhu, Q., Xue, Y.: BEMD image fusion based on PCNN and compressed sensing. *Soft. Comput.* **23**(20), 10045–10054 (2018). <https://doi.org/10.1007/s00500-018-3560-8>
8. Chen, Y.Q., Zhang, L.N., Zhao, B.B.: Identification of the anomaly component using BEMD combined with PCA from element concentrations in the Tengchong tin belt. *SW China. Geosci. Front.* **10**(04), 1562–1576 (2019)
9. An, F.P., Liu, Z.W.: Bi-dimensional empirical mode decomposition (BEMD) algorithm based on particle swarm optimization-fractal interpolation. *Multimedia Tools Appl.* **78**(12), 17239–17264 (2019)
10. Ma, X., Zhou, X., An, F.: Bi-dimensional empirical mode decomposition (BEMD) and the stopping criterion based on the number and change of extreme points. *J. Ambient. Intell. Humaniz. Comput.* **11**(2), 623–633 (2018). <https://doi.org/10.1007/s12652-018-0955-4>
11. Yan, T., Zhou, C.: The research of improving PCA recognition rate of palmprints with BEMD. *CAAI Trans. Intell. Syst.* **8**(4), 377–380 (2013)
12. Huang, D.S.: Radial basis probabilistic neural networks: model and application. *Int. J. Pattern Recognit. Artif. Intell.* **13**(7), 1083–1101 (1999)



Palmpprint Feature Extraction Utilizing WTA-ICA in Contourlet Domain

Li Shang¹(✉), Yuze Zhang², and Zhan-li Sun³

¹ Department of Communication Technology, College of Electronic Information Engineering, Suzhou Vocational University, Suzhou 215104, Jiangsu, China
s10930@jssvc.edu.cn

² Department of Civil, Environmental and Geomatic Engineering, Institute of Finance and Technology, University College London, London, UK
zctpyz3@ucl.ac.uk

³ Department of Automation, College of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, China

Abstract. Contourlet transform can obtain the better contour of an image and make it sparser in local subspace. While independent component analysis based winner-take-all (WTA-ICA) algorithm can extract efficiently image features and is simpler and faster under high dimensional computational requirements. Therefore, combined the advantages of the two algorithms, a new palmpprint feature extraction method utilizing WTA-ICA in contourlet transform domain is discussed in this paper. First, each test image selected from PolyU palmpprint database is pre-processed by using contourlet transform to obtain low frequency and high frequency sub-band images in given layers, and high frequency sub- band images are denoised by the wavelet method. Then the WTA-ICA algorithm is used to train the low and high frequency sub-bands to obtain the low and high frequency features. Further, considered feature fusion method for the low and high features as well as palmpprint original WTA-ICA features, the palmpprint feature extraction task can be well realized.

Keywords: Contourlet transform · Independent component analysis (ICA) · Win-take-all (WTA) · Feature extraction · Feature fusion

1 Introduction

Biometrics-based personal identification is regarded as an efficient method for automatically recognizing a person's identity with a higher confidence [1–3]. While palmpprint verification is just one of the emerging technologies because palmpprint features are stable and unchanged throughout an individual's life [4]. More recently, more and more new palmpprint feature extraction approaches have been explored, such as eigenpalm [2], Gabor filters [3], Fourier transform [4], wavelets transform [5], principal component analysis (PCA), independent component analysis (ICA) [6] and so on. In these methods above-mentioned, ICA methods can be performed well on palmpprint images to separate the important information contained in the high-order relationships and be used

widely in palmprint features extraction [4, 7, 8]. So far, many ICA algorithms behaving different cost functions have been developed [9–11]. In these algorithms, winner-take-all based ICA (WTA-ICA) is simpler and faster under high dimensional computational requirements, and can ensure the maximal sparsity of feature coefficients, so it is used usually to extract palmprint features. Otherwise, in order to ensure the sparseness and the high frequency details of an image, the contourlet transform is also commonly used in the image preprocessing [4]. Therefore, considered the advantages of the both, a new palmprint feature extraction method is proposed in this paper. In test, Palmprint images are selected from the Hong Kong Polytechnic University (PolyU) database. First of all, each palmprint images selected are preprocessed by contourlet transform method, and the low and high frequency sub-bands are obtained respectively. And the high frequency sub-bands of different layers are fused and denoised by the wavelet method. Next, in the contourlet transform domain, the WTA-ICA algorithm is used to train the low and high frequency sub-band image sets respectively, so the low and high frequency features can be well extracted. Further, for features of low and high frequency sub-bands, combined again original palmprint images' features obtained by WTA-ICA, the weighted average feature fusion technique is considered here. And these fusion features obtained are used as the total palmprint image features in the feature extraction task. Furthermore, used the extracted features, an image restore task can be implemented efficiently. Otherwise, compared with features obtained by WTA-ICA, simulation results show that our image feature extraction method discussed here is better in the palmprint feature extraction task.

2 Contourlet Transform

In existing ICA algorithms, the source signals are required to behave sparseness. However, in image processing, many nature image are not sparse. To be ensure the sparseness of images, contourlet transform is used in this paper. This transform is based on an efficient two dimensional non-separable filter banks and provides a flexible multi-resolution, local and directional approach for image processing [9]. It is better than wavelet or wavelet packet transform in dealing with singularity in two or higher dimensions, especially representing images with smooth contours. It is constructed as a combination of the Laplacian Pyramid and the 2-D directional filter banks (DFB) that can be maximally decimated while achieving perfect reconstruction [9]. And it can be simply generalized as two steps [9]: Firstly, to catch all singular points in an image, the multi-scale decomposition is realized by using the LP filter. Secondly, for the HP sub-band images, the DFB are applied to them in order to make all singular points in the same direction composed into a contourlet. The DFB are designed to capture the high frequency components of images, and are efficiently implemented via an l^- level tree-structured decomposition, which can lead to $2l$ sub-bands with wedge-shaped frequency partitioning as shown in Fig. 1, and each sub- band represents a direction in Fig. 1.

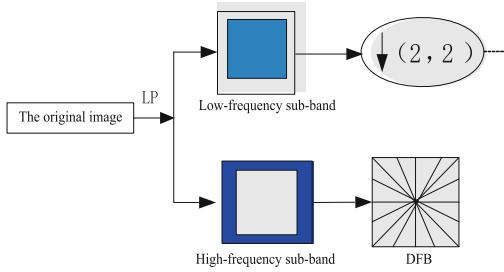


Fig. 1. The original images and the corresponding degenerated images. (a) The original Lena image; (b) Noise version of Lena image with noise level 0.01; (c) The imaging target; (d) MMW image.

3 WTA-ICA Algorithm

For a random vector y behaving zero mean value and unit variance, its sparseness to l^p norm criterion is defined as follows [2]:

$$\text{Sparse}(y) = -E \left\{ \left(\sum_i |y_i^p| \right)^{\frac{1}{p}} \right\}. \quad (1)$$

when $p \rightarrow \infty$, the limit of $\left(\sum_i |y_i^p| \right)^{1/p}$ is written as follows:

$$\lim_{p \rightarrow \infty} \left(\sum_i |y_i^p| \right)^{1/p} = \max_i \{|y_i|\}. \quad (2)$$

Thus, for the l^∞ norm sparseness measure function, the maximization sparseness is equal to optimize the following form:

$$J(y) = E \left\{ \lim_{p \rightarrow \infty} \left(\sum_i |y_i^p| \right)^{1/p} \right\} = E \left\{ \max_i \{|y_i|\} \right\}. \quad (3)$$

In Eq. (3), let $|y_j|$ be replaced with $|y_j|^2 = (w_j^T x)^2$, where x is also a random vector and w_j is the column vector of weighted matrix W , Eq. (3) can be rewritten as Eq. (4):

$$J(W) = E \left\{ \max \left[(w_j^T x)^2 \right] \right\} = \int \max_j \left[(w_j^T x)^2 \right] p(x) dx. \quad (4)$$

For $\lim_{r \rightarrow \infty} \left[\sum_j (w_j^T x)^{2r} \right]^{\frac{1}{r}}$ mathematical convenience, the maximum part $\max_j (w_j^T x)^2$ can be replaced with. Thus, the cost function of WTA-ICA model is obtained as follows:

$$J(W) = \int \lim_{r \rightarrow \infty} \left[\sum_j (w_j^T x)^{2r} \right]^{\frac{1}{r}} p(x) dx = \int \lim_{r \rightarrow \infty} B^{\frac{1}{r}} p(x) dx. \quad (5)$$

In Eq. (5), $B = \sum_j \left(w_j^T x \right)^{2r}$.

The maximum of Eq. (5) can deduce the learning rule of the j th column vector w_j shown in Eq. (6):

$$\frac{\partial J}{\partial w_j} = \int \lim_{r \rightarrow \infty} \frac{\partial B^{\frac{1}{r}}}{\partial w_j} p(x) dx = \int [2 \delta_{cj} (w_c^T x) x] p(x) dx. \quad (6)$$

where parameter δ_{cj} is the Kronecker delta [2]. If $c = j$, then $\delta_{cj} = 1$, otherwise, $\delta_{cj} = 0$. Assumed $(w_j^T x) x$ term with $\|w_j\| = 1$ to be an observation, the goal is to get the mean of this observation, while w_j is estimated incrementally. Considered an amnesic mean, the updating rule of weight vector w_j can be derived as:

$$\begin{cases} w_j(n_j+1) = \alpha(n_j) w_j(n_j) + \beta(n_j) \frac{w_j(n_j)^T x_t}{\|w_j(n_j)\|} x_t \\ \alpha(n) = \frac{n-1-\mu(n)}{n} \\ \beta(n) = \frac{1+\mu(n)}{n} \end{cases}. \quad (7)$$

where n is the n th iteration times, $w_j(n_j)$ is the component vector w_j after the n_j th updating, x_t is the current whitened data input.

4 Experimental Results and Analysis

4.1 Results of Contourlet Transform

Test images used in this paper were selected from the PolyU database. This database includes 600 palmprint images with the size of 128×128 pixels from 100 individuals. First 300 palm images were selected from this database as test images. Supposed the number of decomposition layers to be 2, thus each layer behaves 4 directions. For example, for a palmprint selected randomly, the low and high frequency sub-band images obtained by contourlet transform were shown in Fig. 2. For 300 images selected, after contourlet transform, 300 low pass (LP) sub-band images with the size of 32×32 pixels were obtained and each sub-band image was converted into a column vector, and then the LP test set with the size of 1024×300 was obtained denoted by X_{Low} . At the same time, 1200 high pass (HP) sub-band images of Layer 1 with 64×64 size and those of Layer 2 with 128×128 size were also obtained. In the same way, each high sub-band image was also converted a column vector, thus each layer's HP test set was obtained

denoted respectively by $X1high$ with 4096×1200 size and $X2high$ with 16384×1200 size. Further, $X1high$ and $X2high$ were fused by weighted method and the fused HP test set were denoted by $XHigh$ with the size of 16384×1200 pixels. Thus, the low and high frequency sub-band training sets used to train features were obtained respectively.

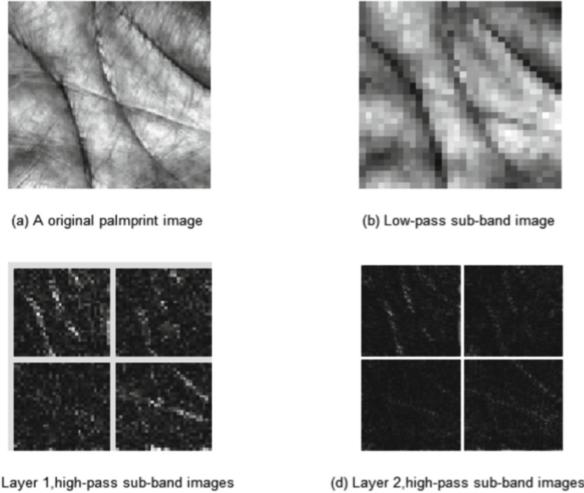


Fig. 2. Several palmprint images selected randomly from the PolyU database.

4.2 Training Features of LP and HP Sub-bands

For the set of LP sub-band images denoted by $XLow$ and the one of HP sub-band image denoted by $XHigh$, each sub-band image was sampled randomly 50000 times by an 8×8 pixel patch, thus the LP and HP training set with the size of 64×50000 were respectively obtained and they were used as the input set of WTA-ICA algorithm, thus the feature vectors of LP and HP sub-band images could be learned, which were shown in Fig. 3. At the same time, the WTA-ICA feature bases of first 300 palm images were also given in Fig. 3. In Fig. 3, the white represent positive pixels, and the gray denotes the zero pixels. Clearly, feature vectors obtained by our method have more distinct sparseness and locality.

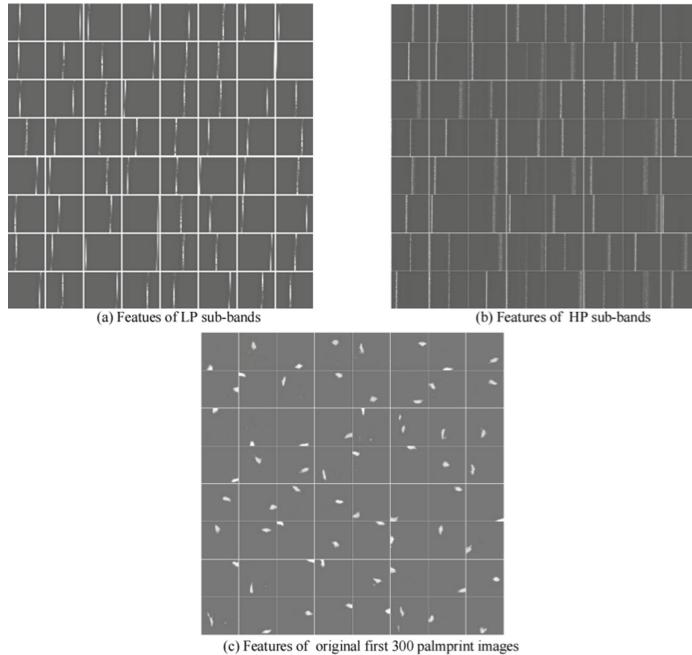


Fig. 3. Original features and ones of LP and HP sub-bands obtained by WTA-ICA algorithm

4.3 Testifying Features

To testify the validity of our feature extraction method, we implemented the image reconstruction task of palmprint images. A test image was randomly selected from the PolyU database. Assume that the number of image patches with 8×8 pixels sampled from this selected image was 10000, 30000, 50000, ..., 100000, etc., respectively. For a palmprint image and its LP image, corresponding to 80000 image patches, the reconstruction results using low frequency features and fusion features were obtained respectively, which were shown in Fig. 4. It is clear to see that three main lines of restored results using low frequency features are very distinct whether it is the original palmprint or its corresponding LP sub-band image, while restored results obtained by fused features have better visual effects. Furthermore, in order to measure the quality of restored images, the signal to noise ratio (SNR) values obtained by using different number of image patches were also given in Table 1. Test results testify that the feature extraction method proposed in this paper is indeed efficient.

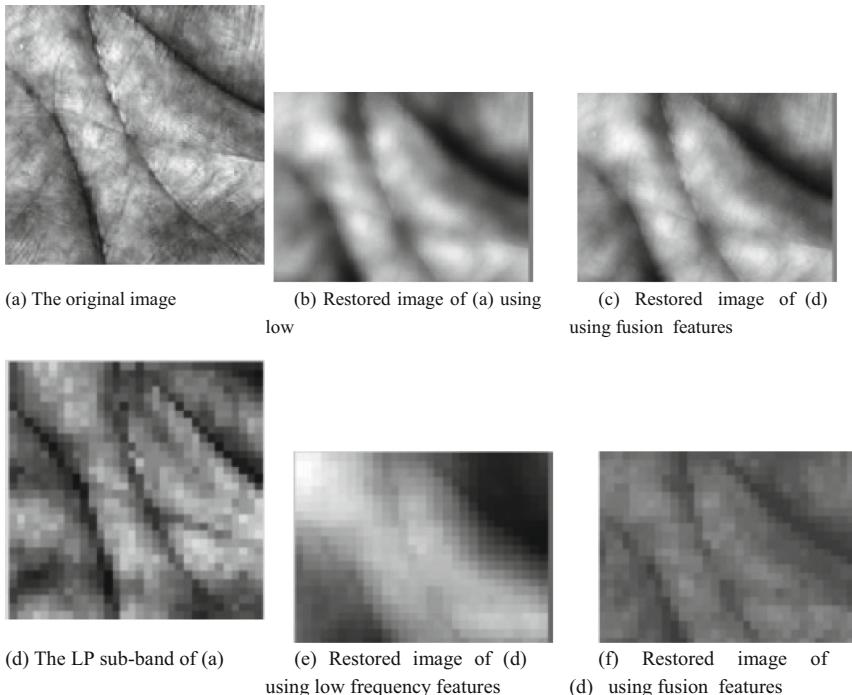


Fig.4. Restored results of a palmprint image and its LP sub-band using different features (80000 image patches).

Table 1. SNR values of different image patches using different methods.

Algorithms	Image patches				
	5000	10000	30000	50000	80000
WTA-ICA in contourlet transform domain	12.582	17.037	21.728	25.527	26.180
WTA-ICA	10.347	15.723	19.823	21.547	24.326

5 Conclusions

In this paper, a novel palmprint image feature extraction method using WTA-ICA in contourlet transform domain is proposed. This method not only consider the low frequency features but also high frequency features. And using the fusion features of the low and high frequency features, a palmprint image randomly selected can be restored well. At the same time, using the image measure criterion of SNR, the efficiency of image reconstruction is proved. Experimental results also show that our feature extraction method proposed in this paper is indeed efficient and useful in practice.

Acknowledgement. This work was supported by the National Nature Science Foundation of China (Grant No. 61972002).

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Trans. Circuits Syst. Video Technol.* **14**(1), 4–40 (2004)
2. Zhang, D.N., Weng, J.: Sparse representation from a winner-take-all neural network. In: Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN 2004), pp. 2209–2214. IEEE Press, New York (2004)
3. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans. Image Process.* **14**(12), 2091–2106 (2006)
4. Bronstein, A.M., Bronstein, M.M., Zibulevsky, M., et al.: Sparse ICA for blind separation of transmitted and reflected images. *Int. J. Imaging Sci. Technol.* **15**(1), 84–91 (2005)
5. Guo, D., Chen, J.: The application of contourlet transform to image denoising. *Control Eng. Inf. Sci.* **15**, 2333–2338 (2011)
6. Yan, Z., Li, Q., Huo, G.: Adaptive image enhancement using nonsubsampled contourlet transform domain histogram matching. *Chin. Opt. Lett.* **A02**, 36–39 (2014)
7. Zhang, C.-J., Nie, H.-H.: An adaptive enhancement method for breast X-ray images based on the nonsubsampled contourlet transform domain and whale optimization algorithm. *Med. Biol. Eng. Comput.* **57**(10), 2245–2263 (2019). <https://doi.org/10.1007/s11517-019-02022-w>
8. Liu, Y.F.: A contourlet-transform based sparse ICA algorithm for blind image separation. *J. Shanghai Univ. (Engl. Ed.)* **11**(5), 464–468 (2007)
9. Hyvärinen, A., Karhunen, J., Oja, E., et al.: Independent Component Analysis. Wiley, New York (1999)
10. Babaie-Zadeh, M., Jutten, C., Mansour, A., et al.: Sparse ICA via cluster-wise PCA. *Neurocomputing* **69**(13–15), 1458–1466 (2006)
11. Lee, H., Battle, A., Raina, R.: Efficient sparse coding algorithms. In: Proceedings of Neural Information Processing Systems, Vancouver, B.C., Canada, pp. 801–808 (2007)



Blockwise Feature-Based Registration of Deformable Medical Images

Su Wai Tun¹(✉), Takashi Komuro¹, and Hajime Nagahara²

¹ Saitama University, Saitama 338-8570, Japan

tun.s.w.748@ms.saitama-u.ac.jp

² Osaka University, Suita 565-0871, Japan

Abstract. We propose a framework for feature-based registration of deformable medical images using a blockwise approach. In our approach, we apply an accelerated-KAZE (AKAZE) feature detector on the initial image frame and input image frames for feature detection, and the detected feature points in the initial image are divided into blocks based on their coordinates. Then, the best feature point in each block of the initial image is picked up by using the response values of the detected features. Tracking of feature points is performed by finding corresponding features between the blockwise features of the initial image frame and all the detected feature points of the current image frame. Our approach has good registration capability even on sparsely textured surfaces such as human organs, which allows our method to be applied for surgery assistance. We demonstrate the effectiveness of our approach using three stereo endoscopic videos.

Keyword: AKAZE feature detector · Deformable registration · Surgery assistance

1 Introduction

In recent years, augmented reality (AR) technology has been applied in various fields. In the medical field, information on operation areas can be provided using AR technology to assist surgery. Since the visualization of the internal structures of organs such as tumors and vessels is important during surgery, AR technology has been applied to enhance the visualization by superimposing tumor or vessel images that are aligned onto the organ.

Image registration is generally categorized into rigid registration and non-rigid (deformable) registration based on the transformation model of an object. The rigid registration allows the mapping between the objects that need to be uniformly rotated and translated, but it cannot change the size or shape of the objects. By contrast, the non-rigid registration allows non-uniform transformation and can map the correspondences between the objects that change their size and shape. Since human organs are non-rigid, changing their size and shape with respiration corresponding, the accuracy of deformable registration plays a crucial role in medical AR.

According to the registration process, registration methods can be divided into two categories: intensity-based methods and feature-based methods. Since feature-based methods are simpler and have lower computational complexity than intensity-based methods, feature-based methods are widely used in medical image registration. On the other hand, the performance of registration can fail when a sufficient number of feature points are not detected [5] or when the number of matched features is small [14, 15].

Some feature-based registration approaches have been proposed to reduce the mismatched features, but they do not evaluate their approaches on the sparse texture surfaces [3, 7]. Some feature matching approaches have been presented [11, 12] to improve the existing methods by finding a large number of correct matched features on deformable surfaces at an increased speed and accuracy. In [11], they can recover and track the features that were lost due to occlusion and eliminate the mismatched features, but their approach has a limitation on poorly textured surfaces and there is a computational limitation in [12], which also cannot retrieve a sufficient set of accurate matches.

In this study, we propose a blockwise approach for feature-based registration of deformable medical images. We apply the AKAZE feature detector on both of initial image frame and input image frames. In the initial image frame, the detected feature points are divided into blocks based on their coordinates and fetch the best feature of each block in terms of the response values of the detected features. Tracking of feature points is performed by finding the correspondences between the blockwise features of the initial image and all the detected features of the input image. Our approach can register the sparsely textured surfaces including human organs. We show the effectiveness of our approach using three stereo endoscopic videos.

2 Related Work

Feature-Based Registration on CT/MR and Histological Serial Section Images

Kajihara et al. [5] described a feature-based non-rigid registration method with a small number of control points for histological serial section images. In their approach, feature detection is performed by accelerated-KAZE (AKAZE), and brute-force matching using the Hamming distance is adopted for the feature matching process. The keypoints are clustered by using their coordinates to determine the local region. Rigid transformation in each cluster is estimated using RANSAC. And rigid transformations are blended to interpolate the transformations at each pixel. Although their method can represent the complex deformation with a small number of control points, it could not perform registration in the image without a sufficient number of feature points. Zhang et al. [14] proposed a hybrid feature detection method for non-rigid registration of lung CT images based on tissue features. The vessel crossing points, vascular endpoints, and tissue boundary points which have high gradients were enough to track the motion of the lung and can be detected by Harris. In this approach, they combined Harris and SIFT to detect blob features since they also used those feature points. Although detected feature points by using their hybrid method were more than those by SIFT, matched feature point pairs by using their method were less than those by SIFT.

Lu et al. [9] improved a linear elastic model for non-rigid medical image registration using the elasticity of the minimum energy as a similarity measure, establishing partial

differential equations to describe the image deformation, and using the finite element method to solve partial differential equations. Their approach is based on the global registration and extracted the feature points of global image registration to generate the irregular triangle grid for defining the region of interest. They showed the robustness of their approach using the 2D CT heart image time series dataset. Although their method can improve the accuracy of registration and enhance image robustness, their method still needs many iterations to converge due to the small shape of the triangle and a large number of triangles.

Zheng et al. [15] proposed a coarse-to-fine registration method based on progressive images and SURF algorithm (PI-SURF). For generating multiple progressive images, the reference image and the floating image are fused. These two images are registered for coarse registration results based on the SURF algorithm and the coarse registration result and the reference image are registered to get the fine image registration. They demonstrated their approach using MR-MR and CT/MR images. In their approach, there are some limitations such as there is time-consuming when the intermediate progressive images are generated and there are mismatching features due to the SURF algorithm.

Feature-Based Registration for Surgery Assistance. Puerto et al. [11] presented a feature matching algorithm called hierarchical multi-affine (HMA) which finds similarities between laparoscopic views. In their method, the detected features are iteratively partitioned into clusters and estimated an affine transformation for each cluster to eliminate incorrect matches from the initial matches. Although the tracked features that were lost are recovered by using affine mapping due to a completed or continuous occlusion or fast camera motion, there are some limitations in their method when the organ or object has poorly textured or when the number of correct matches is very small.

Stoyanov et al. [13] proposed a framework for tissue deformation tracking using a monocular endoscope. Their method was performed by sparse salient features combined with geometric surface parameterization and applied to a phantom heart video sequence to track the motion of an observed fiducial. To compare with ground truth data, the estimated motion of video is compared with the coordinates of the fiducial trajectory of the CT image that was reprojected back into camera space. Their proposed method can estimate the motion of the fiducial, but it cannot handle that their mesh coordinates form with misalignments due to occurring feature matching errors and noise.

Kim et al. [6] proposed a framework for tracking and augmenting a deformable surgical site using shape from shading to recover the 3D shape of the surface and the shape was flattened by using conformal mapping.

Feature detection was performed by using SIFT on that flatten shape and matched the features using Pizarro and Bartoli's feature matching algorithm.

For outlier removal, RANSAC was used. And pseudo-huber norm cost function was used for optimization. Their aim for augmentation is to determine and visualize the boundary regions in the current input frame by matching the features between planer surfaces. Therefore, the boundary of the organ in the planer surface was estimated and the vertices of the boundary are mapped into 3D space by using barycentric interpolation. By using the camera's projection function, these 3D vertices were projected onto the current laparoscopic image for the boundary of the organ. Then, to determine the surgical target, the inlier matches were transformed from planer surface to laparoscopic image,

and affine-MLS was used for smoothly warping the target positions located on the reference image to the current image by feature correspondences. And the locations of the surgical target were marked and overlaid on the top of the original input frame. According to their approach, they can retrieve the surface deformation, but their tracking can fail sometimes due to image blurring and matched features were not found.

Haouchine et al. [2] presented a method for the real-time augmented reality of internal liver structures during surgery. Their approach can locate the in-depth position of the tumors based on partial three-dimensional liver tissue motion using a real-time biomechanical model. To recover the 3D information from the liver surface, they used a stereo endoscope. In their work, Speed-Up Robust Features (SURF) was used to detect salient landmarks in each image pair and tracked by using Lucas-Kanade optical flow since their registration method was a point-to-point registration method. Their method showed good results in terms of surface registration and internal tumors localization. However, the interference may occur due to the outliers when tracking.

3 Proposed Method

In our approach, we apply the AKAZE feature detector for detecting the feature points in the initial image frame and current input image frame. The best feature points among the detected features of the initial image are fetched by using the blockwise approach. Our deformable registration process is performed by tracking the corresponding points between the initial image and the current input image.

3.1 Feature Detection and Blockwise Features

Although there are many methods for feature detection and description, we adopt an accelerated-KAZE (AKAZE) feature detector. Feature detection and description in non-linear scale space are time-consuming due to the high computational load to build the non-linear scale spaces. In the AKAZE feature detector, Fast Explicit Diffusion (FED) scheme embedded in a pyramidal approach can speed up non-linear scale-space construction and its Modified-Local Difference Binary (M-LDB) descriptor that is invariant in rotation and scale can utilize gradient and intensity information from non-linear scale spaces [1]. Therefore, AKAZE features have low computational and descriptor storage demand. The structure of the AKAZE feature contains 2D coordinate positions, a response that describes the strength of the feature, size, class-id, octave that describes the level of scale space, and angle. We use AKAZE in OpenCV with default parameters except for the threshold value. Threshold, one of AKAZE feature detector parameters, allows accepting the feature points. The less the threshold value we set, the more feature points we can get as shown in Fig. 1.

Assume that the initial image frame I_0 and the current image frame I_k are given. Firstly, we apply AKAZE on the initial image frame and divide the detected features into blocks $B = \{ b_0, b_1, b_2, \dots, b_{n-1} \}$ in terms of their coordinates. Then, the advantage of the response field of the AKAZE feature is taken to define the best feature among the detected features in each block of the initial image frame, which we call blockwise

features as illustrated in Fig. 2. The higher value of the response field of a feature, the stronger feature is. Then, we also apply AKAZE on the current image frame for feature detection and description.

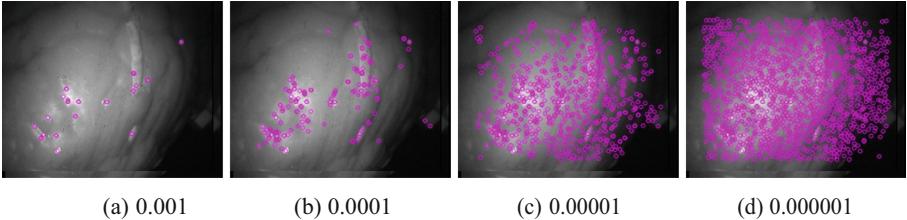


Fig. 1. Detected features with various threshold values.

3.2 Feature Matching and Registration

To track the detected feature points in each input image frame, the corresponding points are detected between the initial image frame and input image frames. Therefore, feature matching is performed by brute-force matching algorithm using Hamming distance to find the correspondence between blockwise features of the initial frame and the detected features of the current image frame.

To eliminate the outlier features from matched results, brute-force matching is combined with the k-Nearest Neighbors (KNN) algorithm and Lowe's ratio test [10]. Lowe's ratio test has simple criteria that determine a good match if the distance ratio of the first closest one and the second closest one is smaller than the threshold value (the default value is 0.5). For outlier removal, we set the threshold value = 0.7 for our cases. The good match that satisfies the threshold value for each image pair contains a feature in a block of the initial image and a corresponding feature of the current frame.

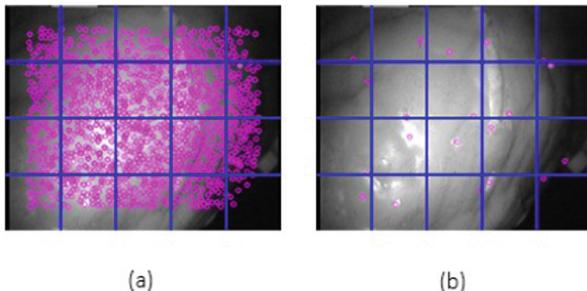


Fig. 2. Blockwise Features Approach: (a) Features in the initial image I_0 are detected by using AKAZE with threshold = 0.000001. The detected features are divided into blocks based on their coordinates and find the best feature in each block using the response value of the feature (b) The best feature in each block of the initial image I_0 .

The coordinates of blockwise features are set as 2D vertex positions of the triangular meshes for the initial image frame while defining the coordinates of the features that correspond to blockwise features as 2D vertex positions of the meshes for the current image frame. We perform the registration by finding features corresponding to blockwise features as shown in Fig. 3. If there is no corresponding feature, we use the corresponding feature of the previous frame image.

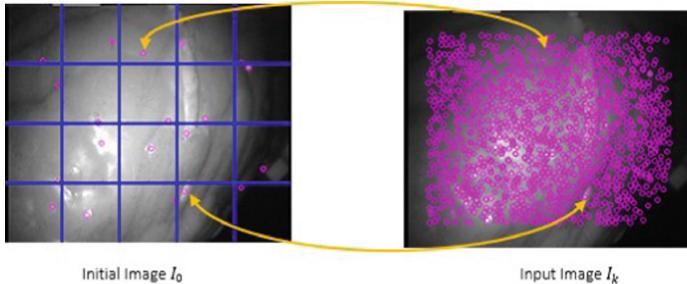


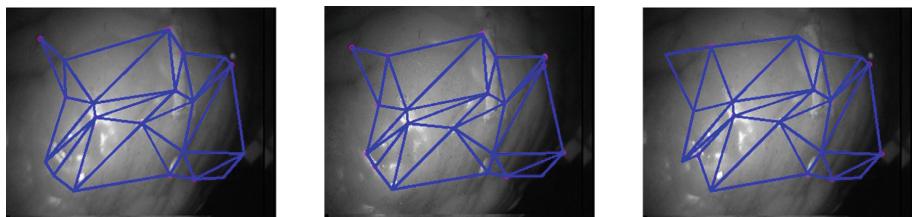
Fig. 3. Finding the corresponding points in the current input image I_k in terms of blockwise features of the initial image I_0 .

To superimpose an image onto the deformable surface in AR, each pixel of the superimposing image in terms of the coordinate system of the initial image frame is mapped onto the current deformed surface by using the bilinear interpolation method.

4 Experimental Results

We conducted an experiment using three endoscopic stereo videos that contained deforming heart and liver from the Hamlyn Centre Laparoscopic / Endoscopic Dataset [8].

We used 194 frame images of Heart-1 video, 100 frames of Heart-2 video, and 49 frames of Liver video. For all videos, ratio test = 0.7 was used to remove the outlier after matching the features from the initial frame image to the input frame images. For the evaluation of registration, the input images were localized onto those in the coordinates of the blockwise features of the initial frame, which we call restored images. The registration and restoration results of Heart-1, Heart-2, and Liver are shown in Fig. 4, Fig. 5, and Fig. 6, respectively. For all videos, we can see the texture in the restored images is similar to that of the initial image.

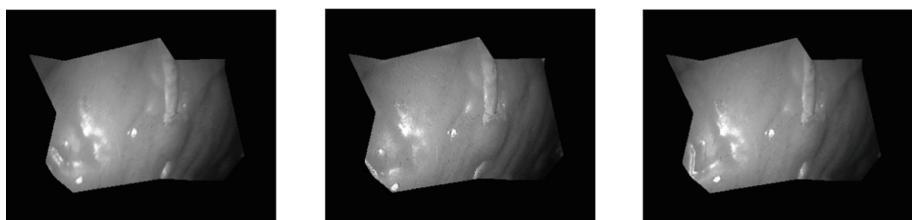


Frame 17

Frame 19

Frame 21

(a)



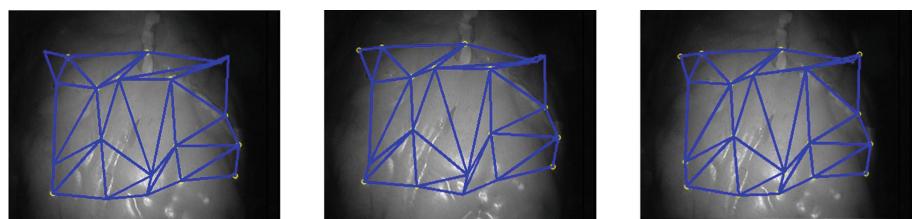
Frame 17

Frame 19

Frame 21

(b)

Fig. 4. The registration and restoration results in each frame of Heart-1. (a) The registration results
(b) The restoration results.

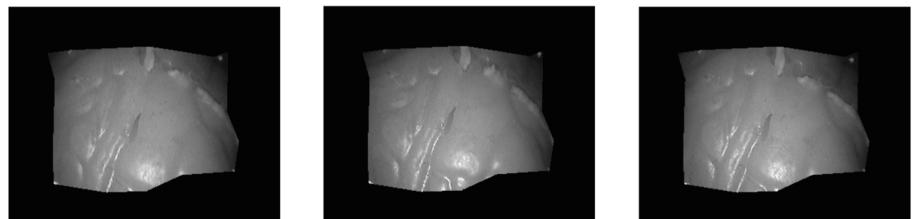


Frame 30

Frame 31

Frame 32

(a)



Frame 30

Frame 31

Frame 32

(b)

Fig. 5. The registration and restoration results in each frame of Heart-2: (a) The registration results
(b) The restoration results.

To evaluate the registration performance in term of texture, we calculated the peak signal to noise ratio (PSNR) between the restored images and the initial image. The PSNR results for all videos are shown in Fig. 7.

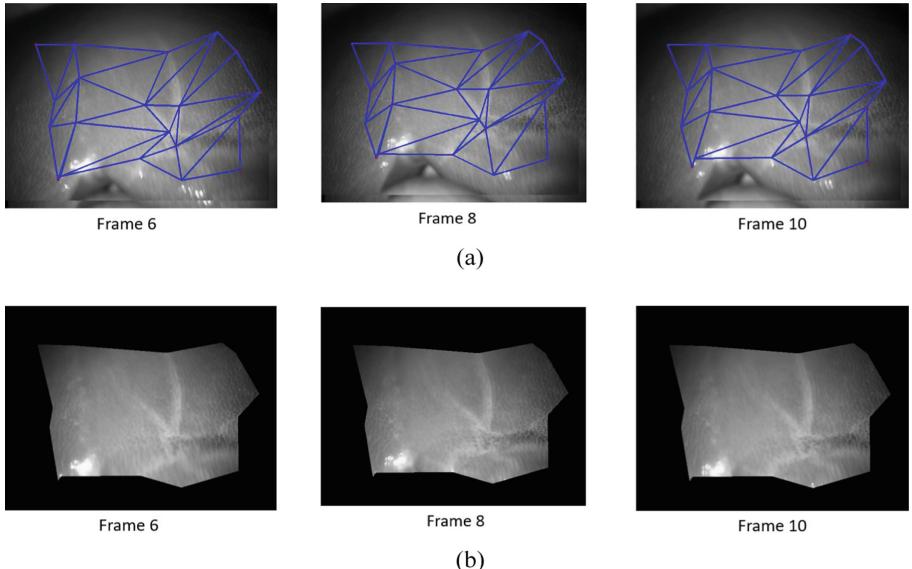


Fig. 6. The registration and restoration results in each frame of Liver: (a) The registration results and (b) The restoration results.

PSNR results of all videos exist in the good performance range. The mean PSNR values of Heart-1, Heart-2, and Liver are 36.6, 36.8, and 36.9, respectively.

As a demonstration for surgery assistance, a hand-drawn tumor image is superimposed onto the deformable surface of each image frame. Figure 8 shows the results of superimposing a tumor image onto an organ. Since our approach can perform on sparsely textured surfaces, we can successfully superimpose a tumor image onto the organs. We can see that the tumor moves along with the movement of the organs.

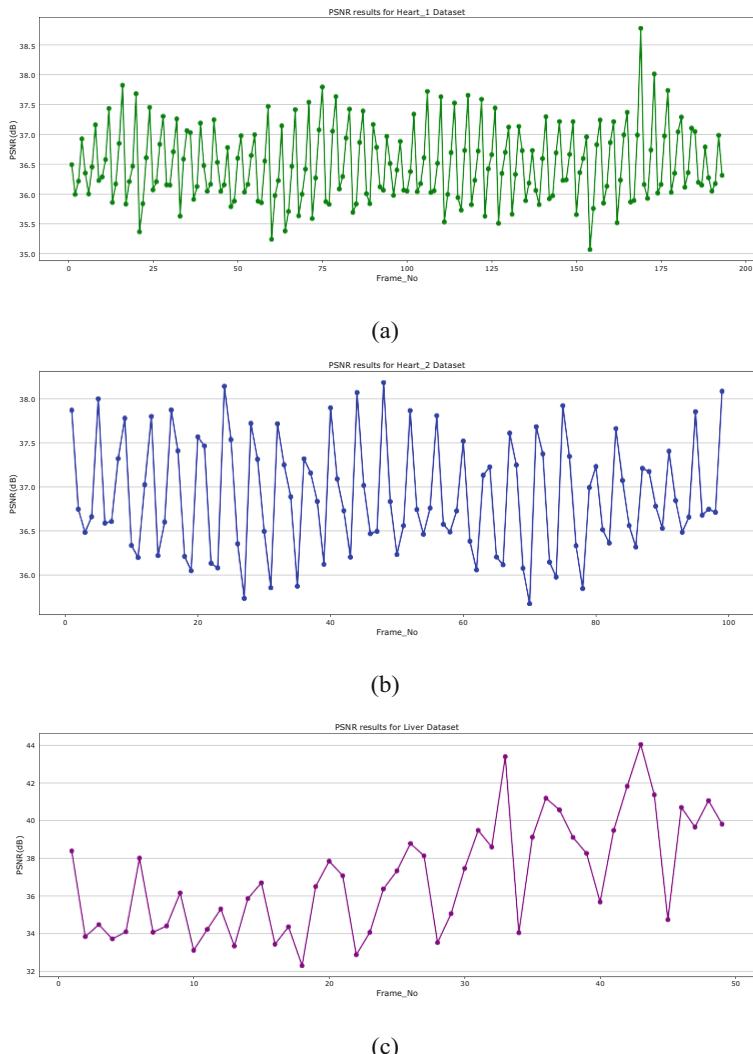


Fig. 7. PSNR results for stereo endoscopic videos in each frame: (a) PSNR results for Heart-1 (b) PSNR results for Heart-2, and (c) PSNR results for Liver.

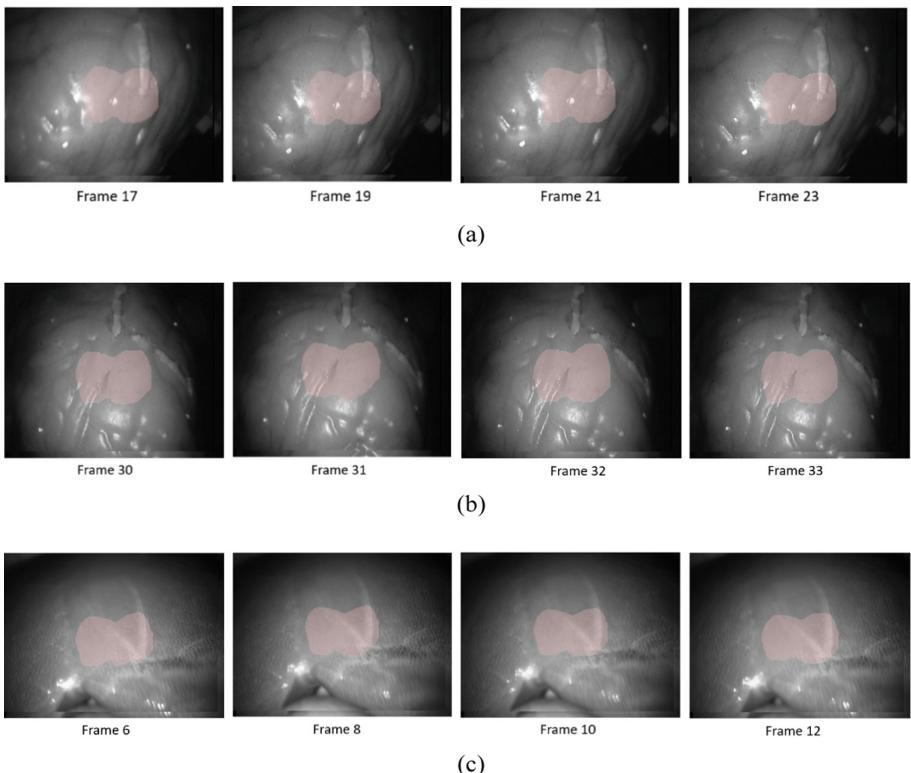


Fig. 8. Superimposing a tumor image on an organ: (a) superimposing a tumor image on each frame of Heart-1, (b) superimposing a tumor image on each frame of Heart-2, and (c) superimposing a tumor image on each frame of Liver

5 Conclusion

We proposed a blockwise feature-based registration framework for deformable medical images. In our approach, the detected feature points in the initial image frame are divided into blocks based on their coordinates, and the blockwise feature points are picked up by using the response values of the detected feature points. Tracking of feature points is performed by finding the corresponding feature points between the blockwise feature points of the initial image frame and all the detected feature points of the current image frame. We presented the effectiveness of our framework by using sparsely textured endoscopic stereo video datasets. The experimental results showed that our framework can be applied even to sparsely textured medical images. In future research, our blockwise framework will be extended for real-time registration.

References

1. Alcantarilla, P.F., Solutions, T.: Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell.* **34**(7), 1281–1298 (2011)
2. Haouchine, N., et al.: Impact of soft tissue heterogeneity on augmented reality for liver surgery. *IEEE Trans. Visual. Comput. Graph.* **21**(5), 584–597 (2014)
3. Hosseini-nejad, Z., Nasri, M.: Image registration based on SIFT features and adaptive RANSAC transform. In: 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 1087–1091 (2016)
4. Jakubovi, A., Velagi, J.: Image feature matching and object detection using brute-force matchers. In: 2018 International Symposium ELMAR, pp. 83–86. IEEE 2018
5. Kajihara, T., et al.: Non-rigid registration of serial section images by blending transforms for 3D reconstruction. *Pattern Recogn.* **96**, 106956 (2019)
6. Kim, J.H., Bartoli, A., Collins, T., Hartley, R.: Tracking by detection for interactive image augmentation in laparoscopy. In: International Workshop on Biomedical Image Registration. Springer, Berlin, Heidelberg, pp. 246–255 (2012). https://doi.org/10.1007/978-3-642-31340-0_26
7. Liao, F., Chen, Y., Chen, Y., Lu, Y.: SAR image registration based on optimized Ransac algorithm with mixed feature extraction. In: IGARSS 2020- 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 1153–1156 (2020)
8. London, I.: Hamlyn Centre laparoscopic/endoscopic video datasets (2019). <http://hamlyn.doc.ic.ac.uk/vision/>. Accessed 15 Jan 2019
9. Lu, X., Ma, H., Zhang, B.: A non-rigid medical image registration method based on improved linear elastic model. *Optik* **123**(20), 1867–1873 (2012)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
11. Puerto-Souza, G.A., Mariottini, G.L.: A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images. *IEEE Trans. Med. Imaging* **32**(7), 1201–1214 (2013)
12. Souza, G.A.P., Adibi, M., Cadeddu, J.A., Mariottini, G.L.: Adaptive multi- a ne (ama) feature-matching algorithm and its application to minimally-invasive surgery images. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2371–2376 (2011)
13. Stoyanov, D., Yang, G.Z.: Soft tissue deformation tracking for robotic assisted minimally invasive surgery. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 254–257 (2009)
14. Zhang, R., Zhou, W., Li, Y., Yu, S., Xie, Y.: Nonrigid registration of lung CT images based on tissue features. *Comput. Math. Methods Med.* **2013**, 7 (2013). Article ID 834192
15. Zheng, Q., Wang, Q., Ba, X., Liu, S., Nan, J., Zhang, S.: A medical image registration method based on progressive images. *Comput. Math. Methods Med.* (2021)



Measuring Shape and Reflectance of Real Objects Using a Handheld Camera

Shwe Yee Win, Zar Zar Tun, Seiji Tsunezaki, and Takashi Komuro^(✉)

Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Sakura-ku,
Saitama 338-8570, Japan

win.s.y.430@ms.saitama-u.ac.jp, komuru@mail.saitama-u.ac.jp

Abstract. In this paper, we propose a system that can measure the shape and reflectance of real objects using a simple apparatus. Our system consists of a handheld camera with a single light source attached, a turntable, and a chessboard with markers. First, using the handheld camera, we acquire images of a target object from multiple viewpoints by capturing a video of the rotating target object placed on the turntable. The shape of the object is reconstructed using the silhouettes of the object seen from multiple viewpoints. For each vertex of the reconstructed three-dimensional shape, the parameters of a reflectance model are estimated using the brightness change in the multiple-viewpoint images. It was confirmed that our system was able to reproduce the appearance of real objects based on the reconstructed 3D shape with reflectance parameters without requiring a large apparatus.

Keywords: Shape reconstruction · Reflectance measurement · Appearance reproduction

1 Introduction

In recent years, with the spread of online shopping, it has become possible to view and purchase products online. However, since users are not able to see the products directly, they may not obtain the correct appearance of the products. We recognize the appearance of a material by seeing the light that comes from a light source and is reflected on the surface. Therefore, to reproduce the appearance of objects, reflectance is important, as well as shape. To fulfill this requirement, a system that can measure both shape and reflectance is required.

There have been measurement systems using multiple cameras and light sources [4, 8, 10]. In these systems, it is possible to measure the shape and reflectance of real objects by observing the light reflected in any direction with respect to the incident light from any direction. However, these systems require a large-scale apparatus, and it also takes a long time to measure the objects.

On the other hand, some researchers have developed methods for measuring the shape and reflectance using only a monocular camera [9, 16]. They captured images of a target object under single incident illumination and estimated the shape and reflectance.

These systems do not require a large measurement apparatus, but the approach used requires a reference object with a shape and reflectance that are similar to those of the target object.

In some systems, multi-view stereo (MVS) has been applied to reconstruct the shape, and reflectance was estimated using the reconstructed shape [12, 18]. The shape and reflectance of object surfaces were estimated from multiple images, assuming that the illumination conditions and camera calibration are known in advance. These systems do not require a large-scale apparatus but may reproduce inaccurate results for objects with highly specular reflectance because feature points are detected between images from different viewpoints.

Based on these studies, we propose a measurement system that can acquire the shape and reflectance of real objects using a simple apparatus. In our system, a handheld camera is used to capture images of a target object viewed from multiple viewpoints while the object is rotating on a turntable. We use a small LED light as a single light source, which is attached to the camera. Images are captured in a dark room where only the LED light is switched on in order to avoid the influence of surrounding light. The target object is placed on a chess board with markers which is placed on the turntable, for acquiring both the camera poses and light directions relative to the target object. The entire shape of the target object is reconstructed using silhouette information in the images captured from multiple viewpoints and corresponding camera poses. In addition, parameters of a reflectance model are estimated from the change in the brightness of each vertex in the reconstructed 3D shape.

2 Related Work

Measuring the shape and reflectance of objects plays an important role for realistic object reconstruction, and many researchers have been active in this field. Existing systems are categorized into three groups:

Shape and Reflectance Measurement Using a Large-Scale Apparatus. To measure the shape and reflectance of objects, some researchers place cameras and light sources in various directions around the target object. Fichet et al. [2] set up an arc-shaped apparatus consisting of a single camera and a rotating arm with multiple lights to acquire photographs under different lighting conditions. Fourier analysis was applied to reconstruct material properties from the sub-sampled signal. Their system can obtain the shape and anisotropic reflectance of the object. Muller et al. [6] used an array of 151 cameras with fixed light sources to record multiple images of a target object. They reconstructed the geometry from the acquired images using visual hull and applied Bidirectional Texture Functions (BTF) to reproduce images of captured objects. Holroyd et al. [14] measured the shape and reflectance of a target object with a setup including a digital camera attached to the arm of a four-axis spherical gantry and acquired images of the target object under a tungsten-halogen light source. Tunwattanapong et al. [3] applied five cameras and a light arc including 105 LED lights to capture images of a target object to reconstruct the shape with reflectance. These systems are effective in providing the shape and reflectance of objects, but they require a large apparatus and long measurement time.

Shape and Reflectance Measurement with a Sample-Based Method. Some researchers estimate shape and reflectance without setting up a large-scale apparatus. Hertzmann et al. [1] proposed a method based on photometric stereo, and they applied a reference object to estimate the shape and reflectance of a target object. In their system, a method based on the photometric stereo approach was used to compute the shape of the target object with reflectance. The reference object has similar shape and reflectance to the target object, and they captured images of the target object and reference object simultaneously under the same lighting condition. Treuille et al. [19] reconstructed the target object with reflectance from the reference object with a known shape and images of the target object. They captured the reference object under the same illumination condition as the target object. In their approach, a voxel-based method was applied to reconstruct the object shape with reflectance. Although these systems do not need a large apparatus, it is necessary to prepare samples of objects with known geometry and reflectance.

Shape and Reflectance Measurement Using a Monocular Camera. Some systems have been proposed for measuring the shape and reflectance of real objects using a monocular camera. Xia et al. [17] proposed a method for shape and reflectance estimation by capturing frames of a rotating object. The shape was reconstructed using a discrete voxel-based approach. After initializing the visual hull and lighting environment, their proposed method was iterated. Their method can estimate shape, light information, and reflectance of the object but materials with high specular reflectance are not suitable for this approach because the accuracy of the reconstructed shape for a sharp specular object is not sufficient to recover sharp lighting. Barron et al. [11] developed a system for estimating the shape and reflectance from a single image of a target object captured with a monocular camera. They proposed the SIRFS model (shape, illumination, and reflectance from shading) which takes a single image of an object as input and reproduces an estimation of the shape, normal, shading, illumination, and reflectance of the object as output. It was possible to use the acquisition system for both shape and reflectance. However, one limitation is that this approach assumes that the target object has a smooth surface and uniform illumination. Giljoo et al. [15] used a smartphone camera with a built-in flashlight to take an image of a target object. This system applied Structure from Motion (SfM) to a group of images captured from multiple viewpoints. Iterative optimization was performed to estimate the reflectance of the surface. As SfM uses feature points for shape reconstruction, detecting feature points between images from different viewpoints does not perform well when the target material has highly specular reflectance.

3 Measurement Apparatus

We use a simple apparatus to measure the shape and reflectance of real objects, as shown in Fig. 1. Using this apparatus, images of a target object from multiple viewpoints

and corresponding camera poses are used to estimate the object shape, and light source positions are used to estimate the reflectance of the target object.

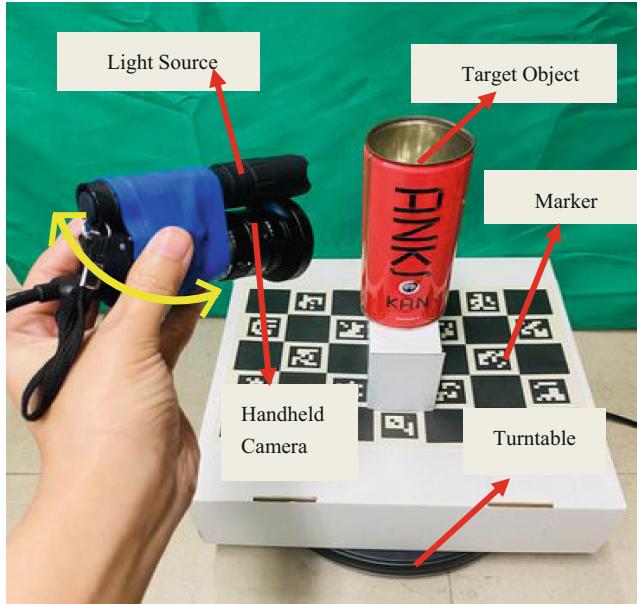


Fig. 1. Measurement setup for our system

The system consists of a handheld camera (Flea3 FL3-U3-13S2C), a turntable, an LED light, and a chessboard with markers. The LED light is used as a single light source, and it is attached to the camera. The camera is moved by hand while being made to face the target object while capturing images. The images are captured in a dark room and only the LED light is turned on. The handheld camera is connected to a PC. The camera captures color frames at a rate of 30 frames per second with a resolution of 1280 x 960 pixels. For the software implementation, the OpenCV library and MATLAB were used, and CloudCompare was used for visualization of the shape reconstruction results. The Unity game engine was used to render reproduced target objects.

4 Shape and Reflectance Reconstruction

4.1 Acquisition of the Camera Pose

We use a ChArUco board [5], which is placed on the turntable together with the target object, to acquire the camera poses relative to the target object, as shown in Fig. 2. The camera pose is obtained as the coordinate transformation from the world coordinate system, which is defined by the markers, to the camera coordinate system.



Fig. 2. Acquiring camera poses using ChArUco board

4.2 3D Shape Reconstruction

The shape of the object is reconstructed from images of the object captured from different viewpoints, as shown in Fig. 3. Visual hull is applied to reconstruct the shape of the object. In visual hull extraction, the common part of the cones obtained by back-projecting is reconstructed as 3D volume data of the object. This method does not require pixel correspondence for each viewpoint and therefore it can be applied to even objects without texture or objects with highly specular reflectance.

To obtain the silhouette image for each viewpoint of the target object, a green background is used. We employ the Moving Least Squares method and Normal Estimation to smooth the surface of the target object.

4.3 Reflectance Estimation

The correspondence between vertices of the reconstructed 3D shape and the color pixels for all captured frames is calculated in order to estimate the reflectance at each vertex. By reprojecting each vertex to all camera images, color pixels that the vertex corresponds to in the images are chosen. The intensity change at the vertex over all the frames is obtained by acquiring the brightness value of the pixel.

The reflectance of the target object is estimated using the reconstructed 3D shape, the obtained intensity change, the distances, and the light angles. The brightness value of each vertex is calculated by applying the following rendering equation:

$$L(t) = \int_{\Omega} \rho_{bd}(w_i, w_o(t)) E(w'_i, t) dw'_i \quad (1)$$

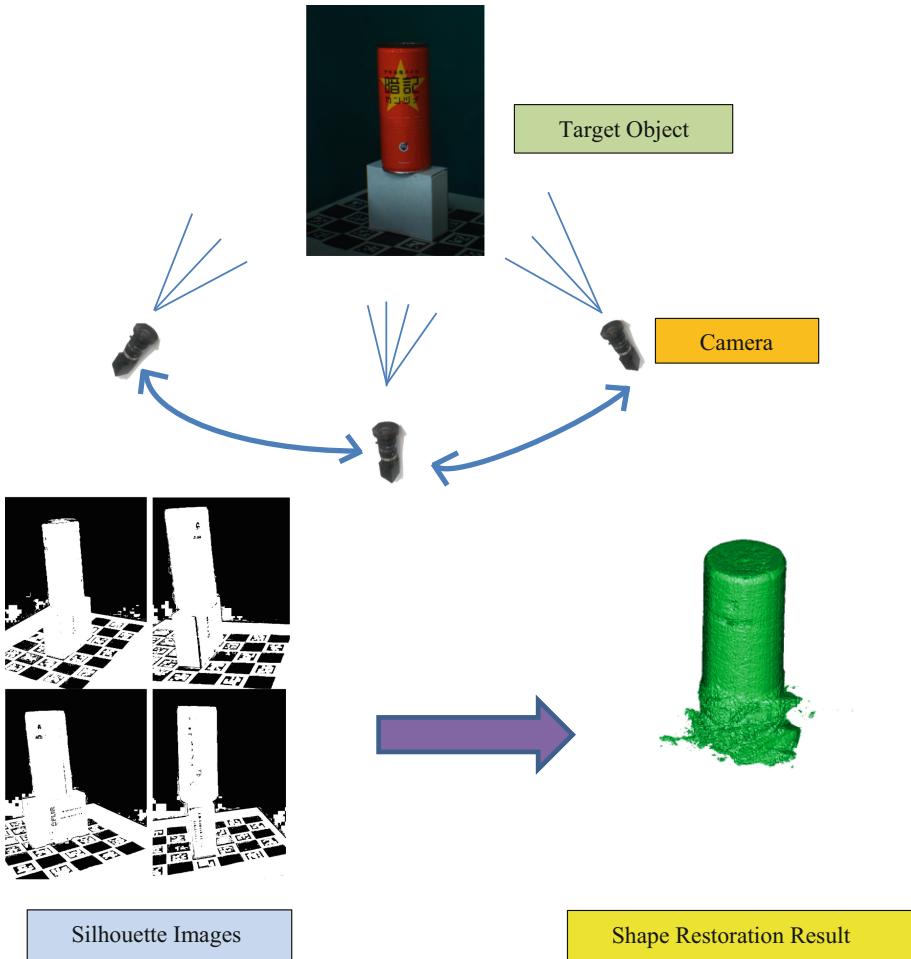


Fig. 3. Shape reconstruction using Visual Hull

where Ω represents the hemisphere whose zenith is the normal direction of the object, $E(w'_i, t)$ is the intensity of light incident from the w'_i direction at time t , $\omega_o(t)$ represents the viewpoint vector at time t and w_i represents the incident vector.

The isotropic Ward model [7] is applied to describe the reflectance of the object. It is the reflectance model that is widely used in computer graphics to fit a measured Bidirectional Reflectance Distribution Function (BRDF) describing how light is scattered at the surface. The isotropic Ward model is expressed as follows:

$$\rho_{bd}(\theta_i, \phi_i, \theta_o, \phi_o) = \frac{\rho_d}{\pi} + \rho_s \cdot \frac{1}{\sqrt{\cos\theta_i \cos\theta_o}} \cdot \frac{1}{4\pi\alpha^2} e^{(-\frac{\tan^2\theta_h}{\alpha^2})} \quad (2)$$

where $\theta_i, \phi_i, \theta_o$, and ϕ_o are the elevation and azimuth angle of the light source vector and the viewpoint vector, respectively, θ_h is the angle between the half vector and the

normal line, and ρ_d , ρ_s and α are the roughness of the diffuse albedo, specular albedo, and specular highlight, respectively. The relationship between each vector used in the reflectance model is shown in Fig. 4. The least squares method is applied to the observed values and estimated the brightness values by the following equation:

$$\arg \min_{\{\rho_d, \rho_s, \alpha\}} \sum_t \|L(t) - I(t)\|^2 \quad (3)$$

In this way, the parameters ρ_d , ρ_s , and α of the reflectance model are estimated for all vertices of the 3D shape of object. The optimization is performed using the Levenberg-Marquardt method, which is one method for solving nonlinear optimization problems.

When applying the nonlinear optimization, it is also necessary to set the initial value. The median value of the obtained brightness values is set as the initial value for diffuse albedo and the value obtained by subtracting the initial value of the diffuse albedo from the maximum luminance value, is set as the initial value for the specular albedo. For the roughness of the specular highlight, a fixed value of 0.5 is set because it is difficult to estimate the initial value.

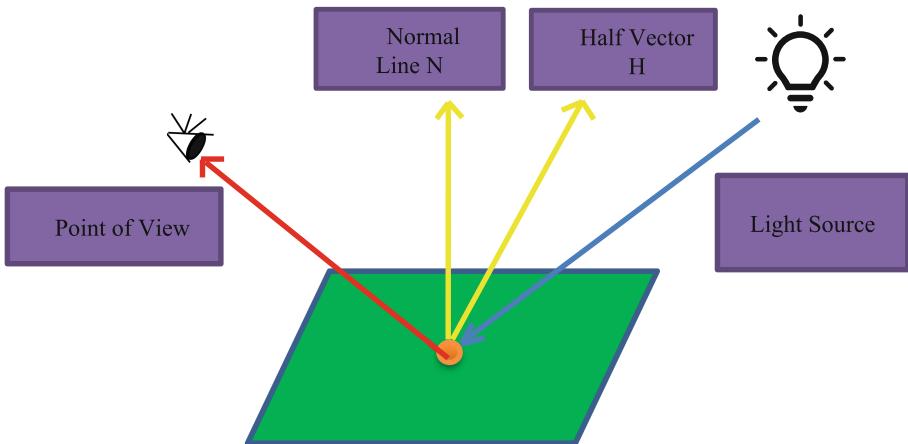


Fig. 4. Vectors used in the reflectance model

5 Experimental Result

In the experiment, real objects were used to measure the shape and reflectance. A red can with strong specular reflection, a black teapot with rough specular reflection and a bowl with no pattern were used, as shown in Fig. 5.

From the results, it was confirmed that even when the objects had strong specular reflection, the visual hull can correctly reconstruct the entire shape of the target object. Moreover, the shape of the object with no texture on its surface was successfully reconstructed, as shown in Fig. 5(c).

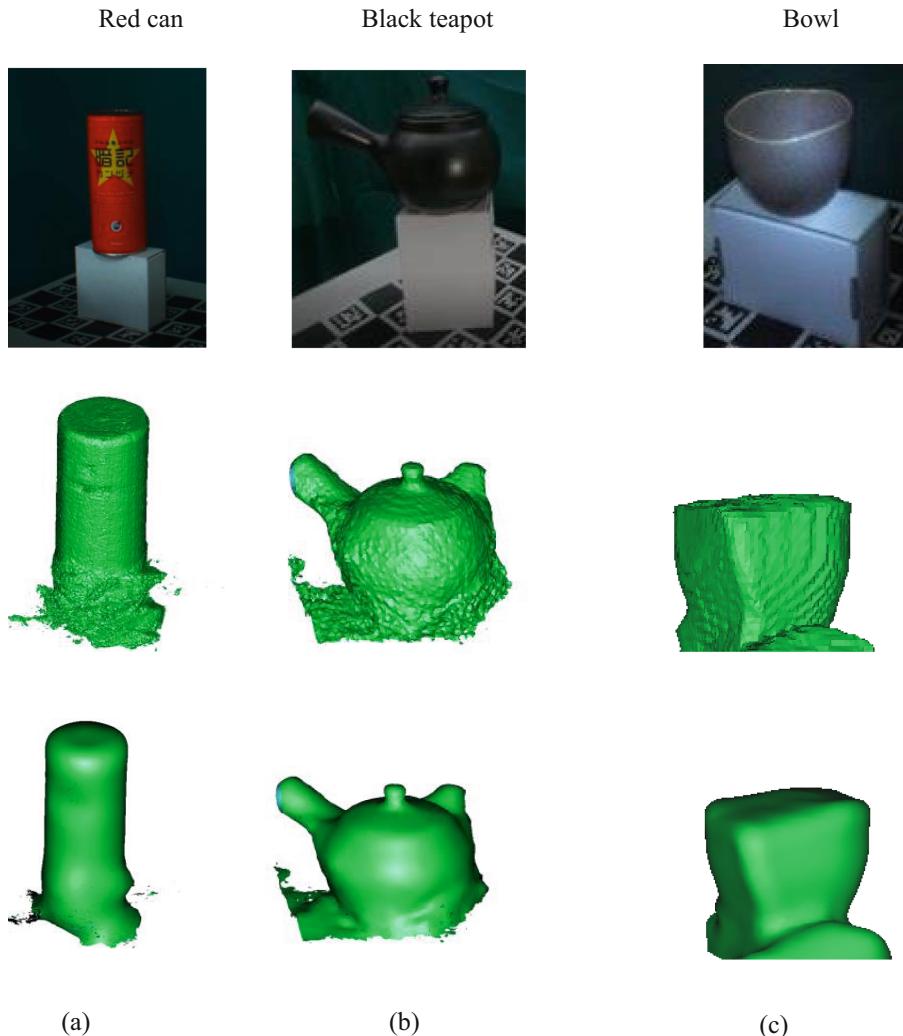


Fig. 5. The shape reconstruction results of different objects by visual hull and the shape results after smoothing. Upper row: input objects, middle row: shape restoration results by visual hull, lower row: shape results after smoothing, (a) red can, (b) black teapot, and (c) bowl.

Figure 6 shows the rendering results of the reconstructed 3D shape of the objects with the reflectance obtained from the estimation of the reflectance model. In Fig. 6(a), the red color of the target object and black characters on the surface were reproduced, and a strong highlight was also observed in the rendering result. In Fig. 6(b), it was confirmed that the same color as the original target object was reproduced. In Fig. 6(c), the color that composes the target object was correctly reproduced with some highlights.

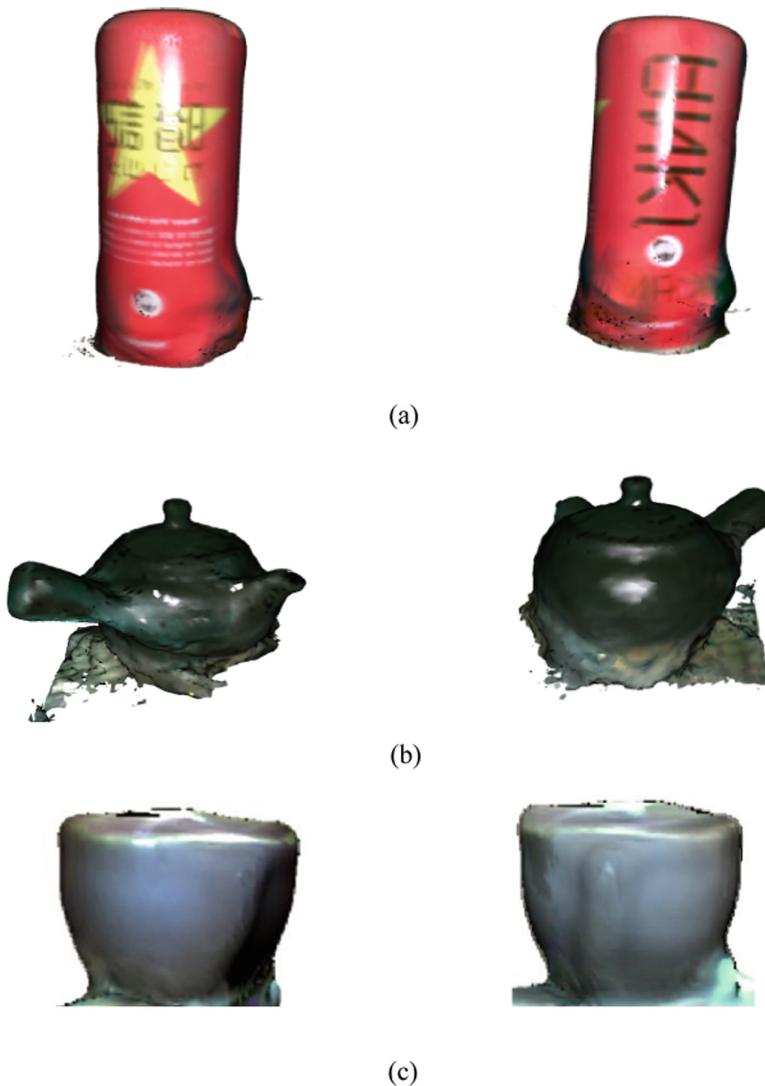


Fig. 6. The rendering results of the reconstructed 3D shapes of objects: (a) red can, (b) black teapot, and (c) bowl.

6 Conclusion and Future Work

In this study, we proposed a system that can reproduce the appearance of real objects with reflectance. Our system does not require a large space and it does not need a long acquisition time. Our system has the limitation that complex-shaped objects and transparent objects are not supported. In the future, we will extend our system for application to a health monitoring system for evaluating edema. Change in edema condition can be visualized by using 3D shapes of the leg. Moreover, skin reflectance is important to

assess the status of the edema. Therefore, we aim to reconstruct the shape of the human leg with reflectance for evaluating edema.

References

1. Hertzmann, A., Seitz, S.M.: Shape and materials by example: a photometric stereo approach. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003) (2003)
2. Fichet, A., Sato, I., Holzschuch, N.: Capturing spatially varying anisotropic reflectance parameters using Fourier analysis. In: Proceedings of Graphics Interface, pp. 65–73 (2016)
3. Tunwattanapong, B., et al.: Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Trans. Graph.* **32**(4), 109 (2013)
4. Schwartz, C., Sarlette, R., Weinmann, M., Klein, R.: DOME II: a parallelized BTF acquisition system. In: Proceedings of the Eurographics 2013 Workshop on Material Appearance Modeling: Issues and Acquisition (MAM 2013), pp. 25–31 (2013)
5. Romero-Ramirez, F.J., Munoz-Salinas, R., Carnicer, R.M.: Speeded up detection of squared fiducial markers. *Image Vision Comput.* **76**, 38–47 (2018)
6. Muller, G., Bendels, G.H., Klein, R.: Rapid synchronous acquisition of geometry and appearance of cultural heritage artefacts. In: Proceedings of the 6th International conference on Virtual Reality, Archaeology, and Intelligent Cultural Heritage (VAST 2005), pp. 13–20 (2005)
7. Ward, G.J.: Measuring and modeling anisotropic reflection. In: Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1992), pp. 265–272 (1992)
8. Li, H., Sing, C.F., Kenneth, T.E., Westin, S.H.: Automated three-axis gonioreflectometer for computer graphics applications. *Optic. Eng.* **45**(4), 1–11 (2005)
9. Riviere, J., Peers, P., Ghosh, A.: Mobile surface reflectometry. In: Computer Graphics Forum, pp. 191–202 (2016)
10. Filip, J., Vavra, R., Haindl, M., Zid, P., Krupicka, M., Havran, V.: Brdf slices: accurate adaptive anisotropic appearance acquisition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 55–69 (2013)
11. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Intell.* **37**(8), 1670–1687 (2015)
12. Yoon, K.-J., Prados, E., Sturm, P.: Joint estimation of shape and reflectance using multiple images with known illumination conditions. *Int. J. Comput. Vision* **86**, 192–210 (2010)
13. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **11**(2), 431–441 (1963)
14. Holroyd, M., Lawrence, J., Zickler, T.: A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. *ACM Trans. Graph.* **29**(4), 1–2 (2010)
15. Giljoo, N., Diego, G., Min, H.K.: Practical svbrdf acquisition of 3d objects with unstructured flash photography. In: Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2018), pp. 267:1–12 (2018)
16. Peiran, R., Wang, J., John, S., Xin, T., Guo, B.: Pocket reflectometry. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2011), pp. 45:1–45:10 (2011)
17. Xia, R., Dong, Y., Peers, P., Tong, X.: Recovering shape and spatially varying surface reflectance under unknown illumination. *ACM Trans. Graph. (TOG)* **35**(6), 1–12 (2016)

18. Yu, T., Xu, N., Ahuja, N.: Recovering shape and reflectance model of non-lambertian objects from multiple views. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 226–233 (2004)
19. Treuille, A., Hertzmann, A., Seitz, S.M.: Example-based stereo with general BRDFs. In: Pajdla, T., Matas, J. (eds.) Computer Vision - ECCV 2004. LNCS, vol. 3022, pp. 457–469. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24671-8_36



Image-to-Video Translation Using a VAE-GAN with Refinement Network

Shengli Wang¹, Mulin Xieshi², Zhangpeng Zhou¹, Xiang Zhang², Xujie Liu², Zeyi Tang², Jianbing Xiahou³, Pingyuan Lin³, Xuexin Xu³, and Yuxing Dai^{3(✉)}

¹ Maintenance Company of State Grid Power Company in Gansu Province, Lanzhou
Gansu 730000, China

² State Grid Info-Telecom Great Power Science and Technology Co., Ltd., Fuzhou 350000,
China

³ School of Informatics, Xiamen University, Xiamen 361005, China
yuxing@stu.xmu.edu.cn

Abstract. With the development of deep learning technology, various techniques for image processing have emerged in the field of computer vision in recent years, and have excellent performance in a variety of application scenarios.

In contrast to the prediction task of predicting video with multiple consecutive frames before and after the input to predict the missing images in the middle, the task of image-to-video generation proposed in this paper does not require multiple consecutive frames, but rather the directional content generation of images by inputting the first frame image with the embedding vector of motion features, and to address some of the existing problems, this paper innovates the network architecture to solve the generated video problems such as incoherence, frame loss and blurring.

For multiple image-to-video translation tasks, we propose a VAE-RGAN network with a further refinement network. We add a refinement network and use new identity matching loss and connected feature matching loss to eliminate VAE and GAN's respective shortcomings and enhance the visual quality of the generated videos. Weizmann datasets have been the subject of a wide range of qualitative and quantitative experiments. We draw the following conclusions from this empirical study: (1) Compared with state-of-the-art approaches, our approach (VAE-RGAN) exhibits significant improvements in generative capability; (2) Experiments shows that our designed VAE-RGAN structure achieves better results and the refinement network significantly improves the problems of a blur.

Keywords: Video generation · Variational autoencoder · Generative Adversarial Network · Refinement network

1 Introduction

The traditional approach regarding the predictive generation of image-to-video, mainly using VAE (Variational Auto-Encoder), is to first input a single frame into the encoder, use a multi-level convolutional neural network to extract and encode the features of the

image, and then reduce it by means of a decoder, i.e. inverse multi-layer convolution, as a way to learn the high latitude features of the single-frame through low subtle parameters, as a way to reach the randomness of motion is modeled, using MES (Mean Square Error) as a loss function to minimize pixel-level loss. This results in the generation of future frames that have some correlation with the previous frame, but the use of MSE, which is essentially an averaging operation on the pixel points of each frame, leads to a lack of clarity of the individual frames in the generated prediction video. Later, with the introduction of GAN (Generative Adversarial Network) technology, people started to experiment with the use of GAN technology for video prediction, which is based on the theory of adversarial gaming through a combination of generators and discriminators, and can generate the desired high-dimensional image content after extensive training and learning. However, due to the limitations of GAN with gradient instability and model collapse-prone, the task of video generation can lead to the generation of future frames that tend to contain too much information from the first frame, resulting in a video motion state that is not obvious and poorly expressed.

Here we present it in three parts:

Video Generation Based on VAE. Recently, the Variational Autoencoder (VAE) has proved to be an essential method for video generation. For example, a VAE was utilized to model the uncertainty of video according to future motion [1, 21, 24]. Generally, the VAE uses the standard loss function, e.g., Mean Square Error (MSE), which causes undesirable motion blur in the generation of video. Currently, two kinds of approaches have been explored to address the problem of blurry video. The first is to use optical flow as the input of VAE to obtain a more stable motion trajectory at the first stage and then to predict the future frames at the second stage [9]. However, this method needs the optical flow to be pre-computed in a way that should well represent the motion change between frames. Actually, optical flow is not always well-aligned between frames in challenging imaging environments. To some extent, misaligned optical flow adds noise to the VAE during the training process. Thus, the generated VAE model may not avoid the blur problem in prediction mode. Furthermore, Pan et.al. presented an alternative way to increase the robustness of the optical flow in an unsupervised manner [14]. However, this work still needs accurate semantic label maps, and it is difficult to obtain semantic label maps. An alternative is to use the beginning and end frames as input for VAE to generate realistic video [23]. Usually, the end frame is unknown in real-world applications, and this will severely limit the potential practical utility of this work.

Video Generation Based on GAN. The GAN has been used in video-related generation tasks and is highly effective in image/video generation applications. For example, several existing video-generation methods [11, 22, 25] have utilized a GAN with multiple frames to successfully predict the next frame or set of frames.

Recently, the work in [20] has assumed that video could be considered as the composite of dynamic objects (so-called foreground) and static invariants (namely background). Based on this assumption, Vondrick et.al. [20] used a neural network to generate the foreground and background separately from the same input latent variable. Also, Saito et.al. proposed a temporal generator based on a single latent variable to generate frames [16]. However, both of these methods [16, 20] suppose that each video can be generated from

a single latent variable. To relax this constraint, Tulyakov et.al. decomposes one short video segment into latent content and motion subspaces [19]. That is, for a single video segment, only the motion latent variable changes, while the content variable does not. Recently, several works have continued to utilize a GAN-based model to solve specific image-to-video translation problems. For example, facial image-to-video translation has been performed by feeding in an action variable together with a neutral facial image [4, 17]. Finally, Nam et.al. have proposed a multi-domain training scheme to generate time-lapse videos [13].

Fusion of VAE and GAN. In [10] and [26], a VAE-GAN model has been applied to text-to-video and text-to-image tasks. Both methods use the VAE to stabilize the training process and also increase the diversity of the sample. They further use a GAN to refine the output of the VAE.

Bao et.al. have proposed a conditional VAE with GAN, namely the CVAE-GAN, to generate fine-grained images from the latent variable, and have subsequently used both a mean feature matching loss and the pairwise feature matching loss to stabilize the training procedure [2]. However, the CVAE-GAN only performs the image generation task. The work in [8] is closely related to [2, 26], but uses the VAE-GAN structure to perform video prediction using two consecutive frames as input. Additionally, Lee et.al. [8] view the GAN as the decoder of a VAE, and so its pixel-level loss affects the generated videos but the blur problem still exists.

In summary, our contributions are three-fold:

1. In this paper, we propose ‘VAE-RGAN’ as the framework for performing multiple image-to-video translation tasks without requiring access to either the first frame of the video or the target category in a coarse-to-fine manner.
2. A video-to-image model is designed with two loss functions to optimize the quality of generated videos and stabilize the training of the whole framework.
3. Comparing the proposed model to state-of-the-art alternatives, both qualitative and quantitative results indicate its superiority.

2 Proposed Approach

The goal is to produce a high-quality video that is both time-coherent and incorporates rich sources of motion information, given an image I_0 and its video target attribute c . In this method section, we introduce the components of the model and the associated training loss functions. Next, we describe the training process for the original VAE-OGAN and the model with the addition of a refinement network (VAE-RGAN). Our training model is then further developed by designing and describing the loss functions. These loss functions reduce the loss of semantic information in the model and preserve the global structure of the object. The novel model ‘VAE-RGAN’ which incorporates the refinement network, is capable of further improving the video quality. Here the refinement network aims at largely reducing the blurry influence of the VAE pixel loss on the output sample.

The VAE-GAN architecture is compared with VAE and GAN for the image-to-video translation task. The total number of frames in the video is T , thus generating a video

sequence of $V = \{I_0, I_1, \dots, I_T\}$, here we work with the following three components for video generation, namely a) blurry video generation, b) blurry video refinement, and c) a generator loss function. Of these, blurry video generation preserves the stability of the model training and generates a blurred video sequence \hat{V} in the VAE-OGAN and in the VAE-RGAN.

In particular, the process of blurry video generation processes the first frame I_0 of the input video sequence V into a 2D convolution layer E_c to extract its semantic content. Tensors of different sizes indicate the semantic content f_c . With the Gaussian model, we extract the motion information z . The VAE encoder E_m is based on 3D convolution layers. During the training of the GAN component, we sampled the Gaussian model $z \sim p_\theta(z)$, where the variance and mean of the model were 1 and 0 respectively, replacing the video $\$V\$$ input to the encoder E_m to obtain motion information z .

In target generated video, c refers to the attribute of the video and can be represented by one hot vector like $c = [100000]$. And we feed z, f_c and c as inputs into the decoder P , to reconstruct the coarse-grained video \hat{V} in our model. The goal of generating high quality video $\hat{V}^g = \{I_0^g, I_1^g, \dots, I_T^g\}$ is achieved by refining the blurred video to improve the quality in VAE-RGAN structure using refinement network R . For the refinement network R , we adopt the U-Net architecture from [15], but the number of channels per convolution layer is reduced by a factor of four. Unit D represents the discriminator in the model and contains two components. The first is the set of 2D convolution layers D_i used to distinguish a single frame between generated frame and real frame, and the second is the set of 3D convolution layers D_v used to distinguish videos between generated video and real video. The output of D is a binary variable.

2.1 Model Training Methods

The traditional VAE-GAN network, i.e., without a refinement network (VAE-OGAN), achieves image to video conversion, but the drawbacks of VAE and GAN are not optimized with the combination of modules. We proposed a new VAE-GAN network architecture, which includes an additional refine network architecture (VAE-RGAN), and performed qualitative and quantitative tests. After refinement of the refine module architecture, the model largely solves the problems of video blurring, distortion, and frame loss generated by VAE-GAN.

In detail, we provide the model with a static first frame and the corresponding target labels, generate a series of deep features of the video frames and the corresponding target label feature vectors by pre-training the encoder, and then generate the hidden vectors $z_0 \sim z_T$ for each frame of the video by using the LSTM network to generate the feature vectors generated by the encoder. The decoder reconstructs the video by taking all the feature vectors and adding the content information of the video frames to ensure the clarity of the generated video. After that, we use Refine-Net to capture the video frames at different scales in order to increase the coherence between the video frames, and ensure the coherence of the overall video transition through the fusion of features between different scales. This improves the authenticity of the video generated by the generator by using the knowledge of game theory to train the overall model (see Fig. 1).

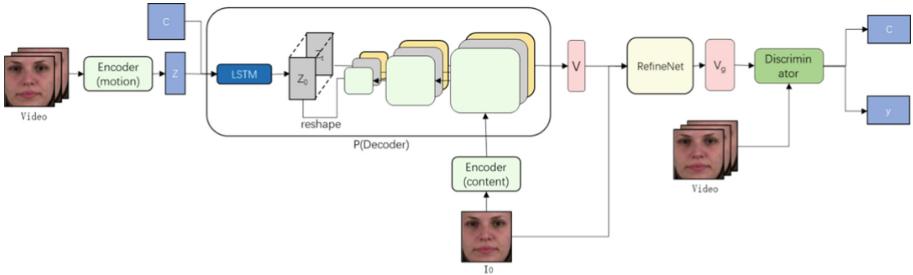


Fig. 1. Illustration of our ‘VAE-RGAN’ decoder structure. For each z , we use the LSTM network to yield $z_0 \sim z_T$, each representing the latent variable of one frame of the generated video. Here, the auxiliary decoder reconstructs the first frame \hat{I}_0 of the video only depending on the input latent variable z_0 . Therefore, the network will not generate videos only depending on the content information f_c .

2.2 Model Architecture

Encoder Architecture. For the behavioral encoder, we used 5 layers of 3-dimensional convolutional layers and 1 layer of 2-dimensional convolutional layers, and a linear equation to extract the potential hidden variables in 512 dimensions. And for the content extraction encoder, we use 4-layer 2-dimensional convolutional layers to extract the features of the input video frames with 256, 128, 64, and 32 units for each channel of extracted features, and the hidden layer vector $z_0 \sim z_T$ in the decoder with a total of 1024 units.

Decoder Architecture. The main decoder mainly uses a two-channel auxiliary decoder architecture based on a channel consisting of four two-dimensional deconvolution modules and two three-dimensional convolution layers. The auxiliary decoder channel has four two-dimensional deconvolution modules. Each 2-D deconvolution module consists of two 2-D convolution layers and one upsampling layer.

Refinement Network. One of the new major modules added to the innovation is the refinement network R, which draws on the U-net network architecture that is popular in the medical image field. In D_V , we use five Conv3D layers, and for D_i , we use four Conv2d layers. The U-net network is a semantic segmentation-based network that has good results in the medical field for image manipulation. It can be seen that the U-net network structure is actually similar to a U-shaped structure. This structure also contains two parts, one of which is a contraction network, while the other is an expanding network. These two structures together form a U-shaped structure that enables effective feature extraction of the input image, while the final convolution operation is designed to map the previously obtained features to the belonging classification.

2.3 Loss Functions

This section is used to introduce the loss functions used by the model. Same as the loss function of the generic VAE-GAN model, the loss function of our VAE-RGAN is

defined as a weighted sum of the five individual losses. The final objective function is summarized as follows. Here, $\lambda_1 = 30$, $\lambda_2 = 5$, $\lambda_3 = 5$ and $\lambda_4 = 5$.

$$L_{Full} = L_{DC} + \lambda_1 L_{VAE} + \lambda_2 L_{Perceptual} + \lambda_3 L_{IFM} + \lambda_4 L_{CFM} \quad (1)$$

Here, L_{DC} is the loss function of the supervisor of VAE-RGAN, mainly to force the generator G to generate the most realistic video possible through game theory, and uses the method of computing cross-entropy to formulate the category labels to specify the target domain of the generated video. The expression of this loss function is shown below.

$$\begin{aligned} L_{DC} = & -E_V[\log D_V(V) - E_{I \sim V}[\log D_i(I)] - E_{z, I_0, c}[\log(1 - D_v(G(z, I_0, c)))] \\ & - E_{\hat{I} \sim G(z, I_0, c)}[\log(1 - D_i(\hat{I}))] - \lambda c \log D_v(c|V)] \end{aligned} \quad (2)$$

And, L_{VAE} is the loss function used to train the VAE component in the VAE-GAN structure, designed to minimize the difference before and after the autoencoder using the KL scatter. It can be written in a generalized form as follows.

$$L_{VAE} = D_{KL}(q_\phi(z|V)||p_\theta(z) + \|\hat{V} - V\|_1 \quad (3)$$

Next, $L_{Perceptual}$ is the perceptual loss who suppresses the difference in output features of the decoder VAE, where V refers to the output features of the real video and \hat{V} refers to the output features of the blurred video.

$$L_{Perceptual} = \sum_i \|\psi_i(\hat{V}) - \psi_i(V)\| \quad (4)$$

Last but not least, L_{IFM} and L_{CFM} are the identity feature matching loss and connected feature matching loss, which are used to improve the consistency of the generated video and the real video content and the quality of the generated video, respectively, and the loss function is shown below.

$$L_{IFM} = \sum_n \left(\frac{1}{2} \sum_t \|\psi_{-1}(I_t^n) - \psi_{-1}(\hat{I}_t^n)\|_1 + L_{FM}(V^n, I_0^n, z) \right) \quad (5)$$

$$\begin{aligned} L_{CFM} = & \frac{1}{2} \sum_t \left\| \psi_{-1}(I_t) - \psi_{-1}(\hat{I}_t) \right\|_2^2 + \frac{1}{2} \|f_{DC}(V) - f_{DC}(G(E_m(V), I_0, c))\|_2^2 \\ & + \frac{1}{2} \sum_t \left\| f_{DI}(I_t) - f_{DI}(\hat{I}_t) \right\|_2^2 \end{aligned} \quad (6)$$

3 Experiments

In the Sect. 3.1, we present the datasets used for the experiments and the corresponding evaluation metrics. And in the Sect. 3.2, we present the experimental details and describe the comparison method as well. Next, in the Sect. 3.3, we compared the qualitative and quantitative experimental results with the two state-of-the-art methods separately. Our method is validated by these results.

3.1 Datasets and Evaluation Metrics

Weizmann Action dataset. The Weizmann Action dataset [3] consists of 90 video sequences showing nine different people, each performing ten natural actions. We pre-process this dataset by following the procedure described in [5, 23] to form the **Weizmann Action-I** dataset. Accordingly, for each video, we divide the sequence of frames into two parts. The first part contains the first 2/3rd of the frames of each video sequence, and we sample several consecutive sequences of 10 frames from the first part and use this as training data. We then consider the remaining 1/3rd of the frame sequence and sample several consecutive sequences of 10 frames and use these as the testing data. In this way, the trained model is aware of the subjects that appear in the testing phase. To evaluate our VAE-GAN structure’s generalization ability, we form the **Weizmann Action-II** dataset. Here the model is aware of the subjects in the training phase, but not in the testing phase. Since the Weizmann Action dataset contains action sequences of nine people, we divide the set of nine people into two parts, we use six people for the training subset and three people for the testing subset. Similar to the **Weizmann Action-I**, we sample several sequences of ten consecutive frames from the two subsets as the training and testing data. As a result, the total number of training samples for the Weizmann Action-I is 2833 dataset and 3385 for the Weizmann Action-II dataset. The total number of testing samples is 810 and 1371 respectively.

Evaluation Metrics. Similar to [1, 23], when generating the videos, PSNR and SSIM metrics are used to evaluate the proposed methods:

1. PSNR uses the formula $PSNR = 10 \cdot \log_{10}(\frac{MAX^2}{MSE})$ to make comparisons of image quality. The formula expresses a mean square error of the maximum value of the color at the image point for both the original image and the processed image. In the experiments, the maximum value MAX_I of the image point color is 255.
2. Through comparing the brightness, contrast, and structure of images, SSIM determines structural similarity. The mean $(2\mu_x\mu_y + \frac{c_1}{\mu_x^2} + \mu_y^2 + c_1)$ is used to estimate the brightness, the standard deviation $(2\sigma_{xy} + \frac{c_2}{\sigma_x^2} + \sigma_y^2 + c_2)$ is used to estimate the contrast, and the covariance $(\sigma_{xy} + \frac{c_3}{\sigma_x\sigma_y} + c_3)$ is used to estimate the structure. Here μ_x and μ_y are the means of the two compared images, σ_x^2 and σ_y^2 are the variance of the two compared images, σ_{xy} is the covariance between the two compared images, c_1 , c_2 and c_3 are constant preventing the denominator from vanishing.

Frechet Inception Distance. Frechet Inception Distance (**FID**) [6] is used for calculating the similarity between a generated image sequence and a target image sequence. As the FID value decreases, the better the model becomes. And we name the variables **FID-ResNeXt** and **FID-I3D** that focus on the image sequence quality as well as the video sequence quality.

3.2 Experimental Setup

In this section, we describe how the training process works.

Training Details. The learning rate and the lambda weights in Equation [公式]. The Adam optimizer [7] is used for our experiments and the hyperparameters β_1 and β_2 are set as 0.5 and 0.999, respectively. In the experiments, the total number of trained batches was set as 100,000 for the Weizmann dataset. The batch size was set as 2 for both datasets. This means that the total number of epochs is about 71 for Weizmann Action-I and 59 for Weizmann Action-II. In our experiments, the learning rate is fixed until the first 80,000 batches for Weizmann, and reduced by a factor of 100 at each epoch. For the more stable gradient obtained from the discriminator, the least-squares loss [12] is implemented for L_{DC} .

Baseline Algorithms. MoCoGAN [19]. In the experiment, we used the conditional image-to-video mode, given the first video frame as input. We used their recommended parameters for training using the Weizmann dataset. In total 12,000 batches are used with batch size 16, which means that the total number of epochs is about 68 for Weizmann Action-I, and 57 for Weizmann Action-II. For the Weizmann dataset, the number of human action categories is 10.

P2PVG [23]. The start and end frames are used to control the video generation process. In our experiments, we aim to observe the visual quality gap between our architectures and P2PVG. The number of epochs is 200, the epoch size is 200, and the batch size is 32.

3.3 Qualitative and Quantitative Results

We now discuss the results appearing in Figs. 2, 3 and Tables 1 and 2 below. Note that the first and end frames are fed into P2PVG. For a fair comparison, the first and end frames are fed into the VAE-RGAN, when we trained Weizmann datasets. For the sake of simplicity, ‘VAE-RGAN’, and ‘VAE-RGAN-p2p’ represent VAE-RGAN with only start frame input, and VAE-RGAN with start and end frames input, respectively.

Qualitative Results. Figures 2 and 3 illustrate the qualitative results on the Weizmann Action-I and Weizmann Action-II datasets, respectively. MoCoGAN suffers from mode collapse during the training stage, which hinders the production of natural videos. By contrast, VAE-RGAN uses the loss function of VAE together with our proposed loss function. This stabilizes the shared VAE decoder and generation network. As a result, VAE-RGAN avoids the mode collapse problem.

Additionally, the visual quality of the video sequences generated by ‘VAE-RGAN’ and ‘VAE-RGAN-p2p’ shows improved resolution compared to the alternative methods. Moreover, they also have improved subjective appearance when compared to the P2PVG model. These results indicate that the ‘VAE-RGAN’ structure can generate higher resolution frames than P2PVG, especially in terms of the background quality.

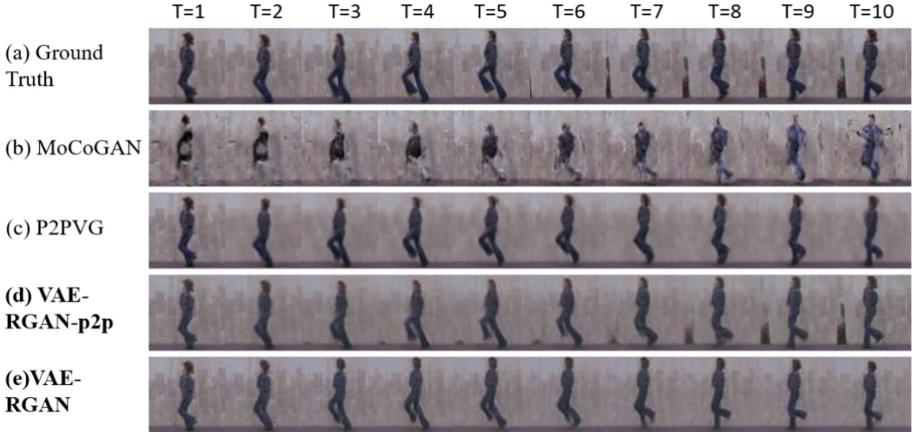


Fig. 2. Comparison of generating videos on Weizmann Action-I. ‘VAE-RGAN’ and ‘VAE-RGAN-p2p’ represent the results of the VAE-GAN structure with refinement network and only the start frame used as input, and the VAE-GAN structure with a refinement network and both start and end frames as input, respectively.

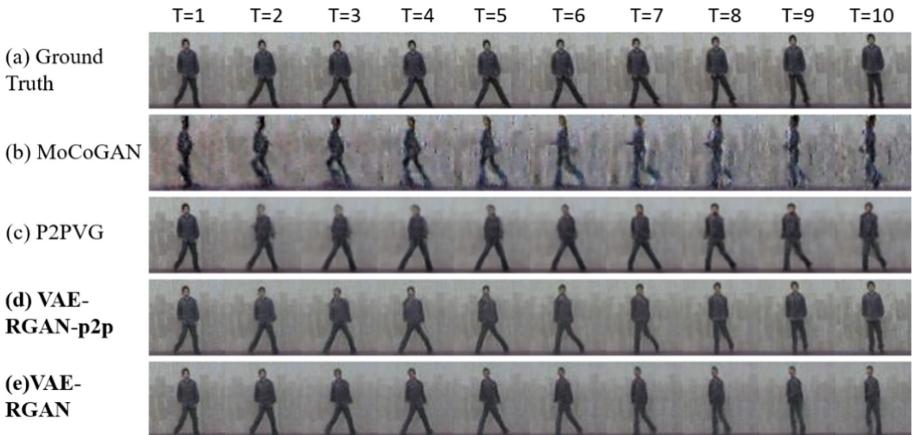


Fig. 3. Comparison of generating videos on the Weizmann Action-II dataset.

Quantitative Results. As seen from Tables 1 and 2, the ‘VAE-RGAN’ outperforms MoCoGAN on the Weizmann Action-I and Weizmann Action-II datasets. Compared with P2PVG, ‘VAE-RGAN’ shows FID-ResNeXt improvements of 0.11 on the Weizmann Action-I dataset and 0.39 of the Weizmann Action-II dataset. This indicates that our proposed method generates higher image quality and more structural similarity to the real video frames than P2PVG. However, our proposed methods perform worse than P2PVG in terms of FID-I3D, PSNR, and SSIM. This is explained by the fact that P2PVG uses the start and end frames as input to improve the consistency of the generated video motions. On the other hand, the P2PVG model used the MSE loss function, which also straightforwardly improves the performance in terms of PSNR. However, the end frame

is usually unknown in real-world applications. As a result, our proposed methods would have more practical value than P2PVG in real-world applications. Finally, for a fair comparison, we further evaluate ‘VAE-RGAN’ by feeding it the start and end frames as input. The results show that two-frame input improves ‘VAE-RGAN’ significantly in terms of FID-I3D, PSNR and SSIM, which improve over those obtained by P2PVG.

Table 1. Comparison of image-to-video generation approaches on Weizmann Action-I dataset.

Model	FID-ResNeXt	FID-I3D	PSNR	SSIM
MoCoGAN	2.71	632.99	19.055	0.400
P2PVG	0.72	162.15	26.925	0.696
VAE-RGAN-p2p	0.45	153.47	27.298	0.788
VAE-RGAN	0.61	191.88	24.102	0.683

Table 2. Comparison of image-to-video generation approaches on Weizmann Action-II dataset.

Model	FID-ResNeXt	FID-I3D	PSNR	SSIM
MoCoGAN	4.97	933.11	18.180	0.362
P2PVG	1.71	292.03	24.560	0.640
VAE-RGAN-p2p	0.65	216.80	25.303	0.739
VAE-RGAN	1.32	329.62	22.018	0.603

Table 3 reports the percentage of selected videos with the best quality on the Weizmann Action-I, and Weizmann Action-II datasets. For a fair comparison, we used VAE-RGAN-p2p for VAE-RGAN on the experiment with the Weizmann datasets. VAE-RGAN outperforms MoCoGAN and P2PVG on the Weizmann Action-II dataset but is still comparable on the Weizmann Action-I dataset. We therefore conclude that our ‘VAE-RGAN’ gives the best subjective video quality.

4 Conclusions

In this paper, by using a starting frame and a label for the target class, we have examined how to generate videos based on a starting frame and label. For the synthesis of high-quality and coherent videos, we present a novel VAE-RGAN structure. For our VAE-RGAN network, we propose a novel loss structure for matching identity features and connected features. In terms of their high-level feature contents, the generated video frames are closer to the real-world video frames by stabilizing the gradient. We have performed extensive experiments using the Weizmann action datasets and VAE-RGAN to demonstrate its superiority. On the image-to-video translation task, our VAE-RGAN model outperforms the competitive state-of-the-art methods based on our comparison

Table 3. Comparisons of human-level performance result in terms of First Best Score between our VAE-RGAN model and baselines on Weizmann Action-I and Weizmann Action-II datasets. “VAE-RGAN or VAE-RGAN-p2p” means that we sum up the scores of “VAE-RGAN” and “VAE-RGAN-p2p”. The best score is in bold.

Model	Weizmann Action-I	Weizmann Action-II
MoCoGAN	1.87%	3.59%
P2PVG	36.39%	16.78%
VAE-RGAN-p2p	33.53%	60.01%
VAE-RGAN	28.21%	19.62%
VAE-RGAN or VAE-RGAN-p2p	61.74%	79.63%

results. The designed architecture and the loss function proposed in the experiments demonstrate that the architecture improves performance while enhancing the architecture as a whole.

References

1. Babaeizadeh, M., et al.: Stochastic variational video prediction. arXiv preprint [arXiv:1710.11252](https://arxiv.org/abs/1710.11252) (2017)
2. Bao, J., et al.: CVAE-GAN: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2745–2754 (2017)
3. Gorelick, L., et al.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2247–2253 (2007)
4. Fan, L., et al.: Controllable image-to-video translation: a case study on facial expression generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01), pp. 3510–3517 (2019)
5. He, J., Lehrmann, A., Marino, J., Mori, G., Sigal, L.: Probabilistic video generation using holistic attribute control. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11209, pp. 466–483. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_28
6. Heusel, M., et al.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Adv. Neural Inf. Process. Syst. **30**, 1–12 (2017)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
8. Lee, A.X., et al.: Stochastic adversarial video prediction. arXiv preprint [arXiv:1804.01523](https://arxiv.org/abs/1804.01523) (2018)
9. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.-H.: Flow-grounded spatial-temporal video prediction from still images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11213, pp. 609–625. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_37
10. Li, Y., et al.: Video generation from text. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32(1) (2018)
11. Liang, X., et al.: Dual motion GAN for future-flow embedded video prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1744–1752 (2017)

12. Mao, X., et al.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2794–2802 (2017)
13. Nam, S., et al.: End-to-end time-lapse video synthesis from a single outdoor image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1409–1418 (2019)
14. Pan, J., et al.: Video generation from single semantic label map. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2019)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2830–2839 (2017)
17. Shen, G., et al.: Facial image-to-video translation by a hidden affine transformation. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2505–2513 (2019)
18. 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)
19. Tulyakov, S., et al.: Mocogan: decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1535 (2018)
20. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. *Adv. Neural Inf. Proces. Syst.* **29**, 1–9 (2016)
21. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: forecasting from static images using variational autoencoders. In: Leibe, Bastian, Matas, Jiri, Sebe, Nicu, Welling, Max (eds.) Computer Vision – ECCV. LNCS, vol. 9911, pp. 835–851. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_51
22. Wang, T.C., et al.: Video-to-video synthesis. arXiv preprint [arXiv:1808.06601](https://arxiv.org/abs/1808.06601) (2018)
23. Wang, T.H., et al.: Point-to-point video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10491–10500 (2019)
24. Xue, T., et al.: Visual dynamics: probabilistic future frame synthesis via cross convolutional networks. *Adv. Neural Inf. Process. Syst.* **29**, 1–9 (2016)
25. Yu, T., et al.: Deep generative video prediction. *Pattern Recogn. Lett.* **110**, 58–65 (2018)
26. Zhang, C., Peng, Y.: Stacking VAE and GAN for context-aware text-to-image generation. In: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1–5. IEEE (2018)



Joint Semantic Segmentation and Object Detection Based on Relational Mask R-CNN

Yanni Zhang¹, Hui Xu¹, Jingxuan Fan¹, Miao Qi^{1,2}, Tao Liu^{2(✉)},
and Jianzhong Wang^{1(✉)}

¹ College of Information Science and Technology, Northeast Normal University,
Changchun 130117, China
wangjz019@nenu.edu.cn

² Changchun Humanities and Sciences College, Changchun 130117, China

Abstract. As fundamental and important problems in computer vision field, semantic segmentation and object detection have made a series of breakthroughs in recent years. Although the existing semantic segmentation and object detection methods have achieved impressive performance in some detection benchmarks, they only focus on local information near the region of objects. However, an image usually contains rich semantic information, including scene context information and dependency information between objects. As a result, ignoring this semantic information will inevitably deteriorate their performance. In this paper, we propose a novel network named joint semantic segmentation and object detection based on relational Mask R-CNN (RM-RCNN) to solve above limitations. By designing the object dependence calculation module (DCM), we can model the relationship information between objects by their geometric and appearance features, so as to improve the accuracy of semantic segmentation and object detection. At the same time, we also design a cross-scale information transmission module (CSITM), which can make the features of different levels transmit information to each other. By using CSITM, our method can effectively retain the useful information and discard the useless information to further improve its performance. Experiments on two benchmark datasets demonstrate the effectiveness of our proposed network.

Keywords: Semantic segmentation · Object detection · Dependence calculation · Information transmission

1 Introduction

With the development of deep learning [1], semantic segmentation and object detection have become active topics in the research field of computer vision and a series of breakthroughs have been made. Nowadays, they have been applied to various fields, such as industry, medical, military and so on.

As we all know, a large number of pixels gathered together form an image, and the task of semantic segmentation is to classify every pixel in the image based on accurate

Supported by Fund of the Jilin Provincial Science and Technology Department (20210101187JC), and the Fundamental Research Funds for the Central Universities (2412020FZ029).

position. Given an image with size of $H \times W \times N$ (H, W represent the height and width of an image, $N = 1$ represents a grayscale image, $N = 3$ represents a RGB image), a matrix of $H \times W \times N$ (C represents the number of categories) can be obtained by using the semantic segmentation algorithm. In this segmentation matrix with the size of $H \times W$, each element has different values (0 or 1) in C channels. Thus, the semantic category of each pixel in the original image can be obtained by the element values at the corresponding position in segmentation matrix. The purpose of object detection is to find the regions of interest in an image and extract their features for classification and regression. In short, the task of object detection is to know both what the object is and where it is located. Combining segmentation with object detection, the dual tasks of recognition and segmentation for each instance in the image can be realized.

For semantic segmentation task, Long et al. [2] proposed a full convolutional network (FCN), which made the segmentation task a big step forward. This network can retain the spatial information and edge information by the fully connected network and jumping connection mechanism. In order to extract more spatial information, SegNet [3] used pooling layer index in the encoder stage and then directly used this index to up-sample features in the decoder stage, which improves the processing speed of the network and facilitates the accuracy of segmentation. Deeplabv1 [4] applied dense conditional random field (CRF) [5] in the segmentation task, which made a better image boundary segmentation performance and improved network performance. He et al. proposed Mask RCNN [6] that adopted pyramid network structure as the backbone network and realized both detection and segmentation tasks.

For object detection task, some methods have begun to explore contextual semantic information to improve their performance [7–16]. For example, Mottaghi et al. proposed a deformable model [7], which utilized the local context and scene-level global context for detection of each candidate box. Torralba et al. [13] proposed a network that punished some objects appearing in unrelated scenes. Shrivastava et al. [16] provided a top-down context to guide regional proposal generation. Although these methods can obtain semantic information about object- or scene-level context through deep learning-based approaches, they made little progress in exploring object-to-object interdependencies. More recently, Chen et al. [17] proposed a new sequential reasoning architecture, which mainly used the interdependence between objects to detect objects sequentially. However, the context-level semantic information is only implicitly considered. The Structural Inference Network [18] not only considered the influence between objects, but also introduced semantic information of scene-level context into the network. This context was described by a kind of structural inference in the network. Experiments showed that this structure can achieve better results. The network proposed by Zhang et al. [19] also considered the relationship between objects and justified the importance of interdependence between objects for detection.

From the above analysis, we can know that most segmentation and object detection algorithms mainly extracted candidate boxes and performed classification regression and pixel classification on these candidate boxes. Although they have achieved impressive performance on some detection benchmarks, these approaches only focused on local information near the object region of interest in the image. However, an image usually contains rich semantic information, including scene context information and dependency

information between objects. Moreover, although some pyramid-based feature extraction networks can extract multi-level features from the input image, they only applied simple operations such as concatenation and jumping connection mechanism when processing intermediate features of different levels, which inevitably affects the performance of the network.

To address the limitations of existing methods and improve the performance of semantic segmentation and object detection, we believe that the dependency between objects should be taken into consideration and modeled at each level of network. At the same time, the information transmission mechanism should also be implemented at different feature levels, so that the information between different levels can be transferred to each other. This kind of design of transmission mechanism can make the network realize the screening of information (keep the useful information, and then filter out useless information). From the experimental results, we can find that the accuracy of segmentation and detection can be greatly promoted by establishing hierarchical dependencies of objects and information transmission mechanism between different layers.

Our main contributions are three-folds:

- (1) In order to enrich the information, we propose an object dependence calculation module (DCM) based on pyramid network. Adding object dependency between hierarchies will play a positive role in both segmentation and detection tasks. Besides, rich detailed information can also help to retain the clear shape and boundary of objects.
- (2) For features of different levels, we propose a cross-scale information transmission module (CSITM), which enables useful information to be transmitted at different levels and ignores useless information.
- (3) Experimental results on MS COCO and PASCAL VOC datasets show that the proposed RM-RCNN can achieve better results than other comparison methods.

2 Related Work

Semantic segmentation: Long et al. [2] used fully convolutional neural network (FCN) to deal with the image semantic segmentation problem. Compared with traditional semantic segmentation algorithm, FCN can make the network get more global information, and also make the network more efficient. This network can segment the input image of any size in an end-to-end, pixel to pixel manner, and then get accurate and fine segmented image. However, FCN does not make good use of high-level features, which makes the network lose some important details. To solve this problem, Chen et al. [4] proposed a semantic segmentation algorithm based on deep convolutional network and fully connected conditional random field (DeeplabV1). DeeplabV1 used dilated convolution to increase the receptive field. Meanwhile, dense conditional random field (Dense CRF) was utilized to integrate global semantic information and recover local details. Zhao et al. [20] proposed a pyramid scene parsing network (PSPNet), which designed a pyramid pooling module composed of four global pooling layers in parallel to realize multi-scale information extraction. The segmentation net-work with the encoder-decoder structure represented by U-Net [21] has achieved good segmentation

results. In the encoder stage, deep semantic feature information can be extracted, and the rich shallow feature information and deep feature information can be fused in the decoder stage to achieve fine segmentation. Pinheiro et al. [22] proposed the Deepmask, which firstly found the candidate boxes of the objects and then identified the corresponding foreground pixel from the candidate boxes. To ensure the ratio of positive and negative samples, the Deepmask specified that each candidate box must contain an object. Dai et al. [23] proposed instance-sensitive fully convolutional network (InstanceFCN), which enhanced the spatial correlation between pixels by designing a dual-branch network structure. He et al. [6] proposed the Mask R-CNN, which realized the dual task of semantic segmentation and object detection by adding a mask branch based on Faster R-CNN. Like Faster R-CNN, the first two branches of this network accomplish classification and regression tasks. In the third branch, each region of interest generates a series of masks, and the network classifies the masks to achieve segmentation task. However, Mask R-CNN neglects the dependence between objects during feature extraction, thus the lack of such semantic information inevitably degrades the accuracy of detection and pixel segmentation to a certain extent.

Object Detection: Nowadays, many researchers have carried out in-depth studies on object detection task, the representative methods include R-CNN [24], Fast R-CNN [25], Faster R-CNN [26] and so on. These methods have some common characteristics: they first generated a series of candidate boxes in the first stage, and then divided these candidate boxes into foreground or background in the second stage. R-CNN was a pioneer deep learning based algorithm for object detection task, which mainly combined several processes such as candidate region selection, feature extraction, feature classification and bounding box regression. The candidate region selection was to select regions of interest by selective search method [27]. Feature extraction was utilized to send the regions of interest obtained in the previous step into the convolutional neural network for feature extraction. Classification refers to the classification of features to their corresponding categories, and bounding box regression refers to the accurate location information. The linear support vector machine (SVM) [28] was applied to achieve classification. Although R-CNN has made great achievement in object detection task, there is still limitation: a mass of candidate boxes needs to be extracted in advance, which will take much time and memory space. To solve this problem, Fast R-CNN [25] was proposed. Fast R-CNN still applied selective search to generate candidate regions. Then, the candidate regions were sent into the convolutional network. The features of any size got by the convolutional network were mapped to each input region of interest through the ROI pooling. Finally, a fully connected network was used to classification and regression. Although the accuracy and running time of this network are greatly improved compared with R-CNN, Fast R-CNN still adopted the selective search strategy to extract candidate boxes, which was time-consuming and failed to meet the needs of real-time applications. Thus, Ren et al. proposed Faster R-CNN [26]. Compared with R-CNN and Fast R-CNN, Faster R-CNN proposed Region Proposal Network (RPN) to achieve the candidate region generation. The core idea of RPN is the use of convolutional neural network to directly generate proposal regions without generating multiple candidate boxes in advance. In essence, RPN simply slides over the last convolution layer to extract feature boxes on the feature map. RPN network is also a full convolutional

network, which can predict the boundary and fraction of the object at the same time with end-to-end training mode. Recently, SSD [29] and YOLO [30] have been proposed for real-time detection with satisfactory accuracy. While these methods worked well with prominent objects in most cases, they cannot effectively deal with small objects by using features that are only relevant to the object itself, so it is important and necessary to take advantage of contextual semantic information.

3 Proposed Method

3.1 Overview

The overall architecture of the proposed RM-RCNN is shown in Fig. 1. It can be seen that RM-RCNN is composed of four parts. The first part is the backbone network based on feature pyramid model, which is employed to extract the feature information of input image. The second part is the hierarchical dependency calculation module, which is used to calculate the interdependence between different objects in the same network layer. The third part is the cross-scale information transmission module between different layers, which is utilized to transfer information between different feature scales. The last part is the feature processing module that is eventually used for classification, bounding box regression and mask classification. We will elaborate on how each part of the network works in the following subsections.

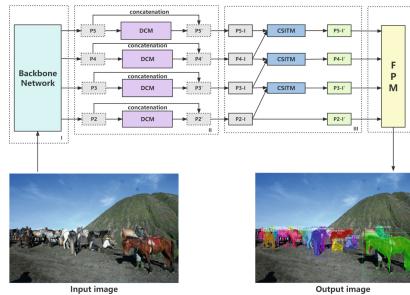


Fig. 1. The architecture of joint semantic segmentation and object detection based on relational mask R-CNN (RM-RCNN).

3.2 Backbone Network

We know that the deep network is powerful to extract the semantic features of images, but its deeper feature map is relatively small and contains little geometric information. This phenomenon has a certain impact on object detection task. Although the shallow features have more geometric information, the semantic information is not rich, which will influence the performance of classification. Therefore, it is necessary to use both deep and shallow features to meet the needs of classification and detection. As we have mentioned in Sect. 2, some classical methods such as Faster R-CNN have improved

detection speed, but they all used single-scale feature maps, which inevitably imposed certain limitations on the detection capability of the network. Thus, feature pyramid model is an ideal backbone network for hierarchically feature extraction. In our RM-RCNN, feature pyramid model is also adopted as our backbone network, and its structure is shown in Fig. 2.

As shown in Fig. 2, let s suppose the input of backbone is an image with the size of $H \times W \times C$. Through four convolution operations (Conv1, Conv2, Conv3 and Conv4) and pooling operations, we can get the feature maps (C_1, C_2, C_3, C_4 and C_5) with the different sizes and channels. At the same time, in order to improve the efficiency of calculation, we use a 1×1 convolution to reduce the dimension of features and obtain features M_2 - M_5 . It is worth noting that C_1 feature is not used in our model because the size of C_1 is too large and will consume a lot of memory during training. Since shallow features (e.g., M_2) contain more geometric information, while deep feature maps (e.g., M_5) contain more semantic information, we adopt a top-down strategy to fuse these features. Specifically, starting from the smallest feature, we apply an up-sampling operation based on bilinear interpolation to gradually get a larger feature. At the same time, in order to retain important details, we also add the feature information obtained in the previous layer to the up-sampling output. In this way, the information of features with different sizes is integrated with each other. After that, we use a 3×3 convolution operation in each layer to reduce the aliasing effect caused by up-sampling. Finally, we get the multi-scale output features (P_2 - P_5) of the input image.

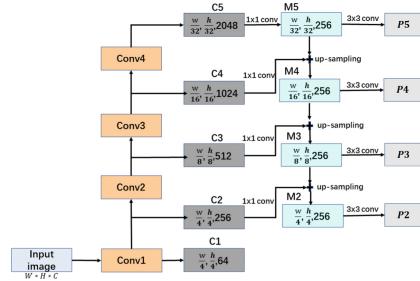


Fig. 2. The architecture of backbone network.

3.3 Dependence Calculation Module (DCM)

Most of the object detection methods based on convolutional neural network identify each object independently, thus ignoring the relationship between objects. Although some methods have modeled the correlation between objects, the dependence between features at different scales is still not been fully considered. To overcome this limitation, we propose to model dependencies between objects in different layers. The process of our DCM is shown in Fig. 3.

Given the features f_i, f_j extracted from two proposals and their coordinate information. First, we calculate the geometric dependence relation R_g between them by fusing coordinate information of different proposals. Here, we take several sets of geometric

coordinates, absolute value, square, logarithm and so on. By Eq. (1) we can get the geometric dependence information R_g .

$$R_g = (w_i, h_i, s_i, w_j, h_j, s_j, |x_i - x_j|, |y_i - y_j|, \sqrt{\frac{x_i - x_j}{w_j}}, \sqrt{\frac{y_i - y_j}{h_j}}, \frac{x_i - x_j}{w_j}, \frac{y_i - y_j}{h_j}, \frac{(x_i - x_j)^2}{w_j^2}, \frac{(y_i - y_j)^2}{h_j^2}, \log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right)) \quad (1)$$

where w , h and s represent the width, length and area of the proposals respectively.

Then we update R_g continuously with W_r . Thus, we can obtain the geometric information f_R by Eq. (2).

$$f_R = W_r * R_g \quad (2)$$

where W_r is the weight that can be learned by network.

After the concatenation operation, we can obtain the feature f_T that integrates the information of different proposals by Eq. (3).

$$f_T = W_t * [f_i, f_j] \quad (3)$$

where W_t is also a weight that can be learned by network.

In order to obtain the weight of dependence between objects, Eq. (4) is applied.

$$Q_g = (\text{relu} \cdot f_R) * (\tanh \cdot f_T) \quad (4)$$

where both relu and \tanh are activation functions. Then, we use the obtained dependency weight Q_g to update the related features, which is expressed by Eq. (5).

$$f = Q_g * f_i \quad (5)$$

Finally, with max pooling operation, the features with dependency information can be got. Thus, by applying the proposed DCM on features of different layers (i.e., P2-P5), we can get the updated features P2-P5 of different layers respectively. In order to keep the previous information, we will fuse input features with these features together, and finally get the output features P5-I, P4-I, P3-I and P2-I of the second part.

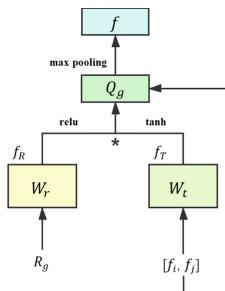


Fig. 3. The architecture of dependence calculation module (DCM) in each scale.

3.4 Cross-Scale Information Transmission Module (CSITM)

In the previous section, we model the interdependence relationship between objects at the same level, and make the feature information of objects richer by DCM. For the pyramid network, the information in different layers may contain some abundance, but how to make the useful information transfer to each other? Based on this thinking, we filter information between different layers by passing on useful information and discarding useless information. Therefore, a cross-scale information transmission module based on LSTM [31] is proposed. The structure of CSITM is shown in Fig. 4.

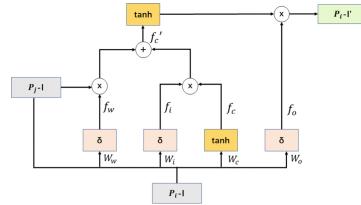


Fig. 4. The architecture of cross-scale information transmission module (CSITM).

Given the input features $(P_i\text{-I}, P_j\text{-I})$ of two adjacent layers. We multiply them by W_w to determine how much information will be forgotten. Through Eq. (6), the degree of forgetting f_w can be got.

$$f_w = (\delta(W_w \cdot [P_i - I, P_j - I]) + b_w) \quad (6)$$

where δ represents the operation of sigmoid, b_w is obtained by network initialization.

Then, to find out how much information can be remembered and passed on, we multiply the input features by W_i . According to Eq. (7), the degree of memory f_i can be obtained.

$$f_i = (\delta(W_i \cdot [P_i - I, P_j - I]) + b_i) \quad (7)$$

where meanings of parameters W_i and b_i are similar to Eq. (6). It is worth noting that we also use a tanh function to obtain the feature f_c for subsequent information selection rather than as a gated signal. The process can be expressed by Eq. (8).

$$f_c = (\tanh(W_c \cdot [P_i - I, P_j - I]) + b_c) \quad (8)$$

In order to get how much information can be output, we use Eq. (9) to model what will be considered output of the current state.

$$f_o = (\delta(W_o \cdot [P_i - I, P_j - I]) + b_o) \quad (9)$$

At last, the entire CSITM process can be represented by the following Equations.

$$f'_c = (P_j - I \cdot f_w + f_i \cdot f_c) \quad (10)$$

$$P_i - I' = (f_o \cdot \tanh(f'_c)) \quad (11)$$

Through Eqs. (10) and (11), we can obtain useful information and discard useless information, which realizes the transmission of information between different layers. This mechanism is different from the simple concatenation and jumping connection mechanism since it can obtain richer information by transmit the features between different layers and make our network be more flexible.

3.5 Feature Processing Module (FPM)

Through feature refining in Sects. 3.3 and 3.4, we can get features that not only contain the dependency relationship of objects in the same layer but also integrate information of different layers. In this section, we will take them as the input of FPM shown in Fig. 5.

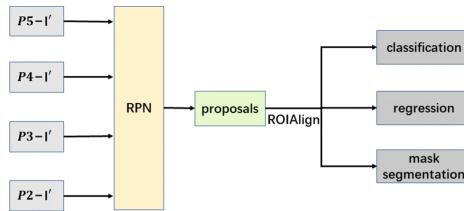


Fig. 5. The architecture of feature processing module (FPM).

According to Fig. 5, we first use the RPN to generate some proposals. Then ROIAlign operation is applied to handle these proposals. The reason why the ROI-pooling operation is not used here is that the network needs to perform two times quantization operations (image coordinates to feature map coordinates and feature map coordinates to region of interest feature coordinates) to obtain the fixed-size feature map. Such two quantization operations will inevitably lead the coordinates of proposal to be floating number rather than integer, which may have little effect on detection task, but seriously deteriorates segmentation task. Therefore, ROIAlign with bilinear interpolation is used to avoid such errors. Through the ROIAlign operation, we can get the fixed-size features for the final classification, regression and mask segmentation tasks.

The total loss of the network can be expressed by the following Eq. (12).

$$L = L_{cls} + L_{reg} + L_{mask} \quad (12)$$

where L_{cls} represents the loss of classification branch, L_{reg} represents the loss of regression branch, L_{mask} represents the loss of mask branch. The definitions of these three loss functions are the same as those in [6].

4 Experiment

In order to verify the performance of the proposed RM-RCNN, we conduct experiments on MS COCO [32] and PASCAL VOC [33] datasets, and compare the results of our method with some other algorithms.

4.1 Experimental Settings and Datasets

Experimental Settings. The experimental hardware environments of RM-RCNN are Intel (R) Core (TM) CPU 3.50 ghz, NVIDIA Tesla V100 16 GB. The operating system is Ubuntu 16.04 and the compilation environment is python3.7. Pytorch1.1 is deployed to build a deep learning framework. During the experiment, our network has no restriction on the size of input image, but we set the definition of IMAGE_MIN_DIM = 1024, which will unify the image to the set size. In the training, we set the initial learning rate as 0.001, momentum as 0.90, weight decay as 0.0001, batch size as 16 and RoIAlign parameter as 0.7.

Experimental Datasets. MS COCO dataset is built by Microsoft, includes 91 classes of objects with 328,000 images and 2.5 million labels. PASCAL VOC dataset is mainly oriented to five tasks: image classification and detection, object detection, motion classification and large-scale object detection. There are 9,963 images with 20 classes of objects.

4.2 Evaluation Criteria

The most commonly used criterion to evaluate the performance of detection algorithms is average accuracy (AP), which comes from precision and recall. The calculation processes of precision and recall are shown in Eqs. (13) and (14).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

where TP represents the number of true positive examples, FP represents the number of fault positive examples, and FN represents the number of fault negative examples. To compare the performance of all object classes, we use the mean AP as the final evaluation indicator.

4.3 Performance Comparisons

We will justify the effectiveness of the proposed network from two sections. The first section compares the performance of our network with some existing algorithms, and the second section utilizes ablation scheme to show the importance and necessity of the components in our network.

Quantitative and Qualitative Evaluation. In Table 1, we compare the proposed method RM-RCNN with the existing algorithms on MS COCO dataset. The comparison algorithms are as follows:

MNC [34]: This method is a classical algorithm for simultaneous detection and instance segmentation, which adopts multi-task mode. On the basis of shared features, the

tasks of adjacent stages in MNC are attached with each other, thus it forms a hierarchical multi-task structure.

FCIS+ OHEM [35]: This method is a fully convolutional end-to-end instance segmentation network with fully shared convolutional representation between two subtasks and between all regions of interests. At the same time, it also realized the dual task of detection and segmentation.

FCIS+++ OHEM [35]: Compared with FCIS algorithm, this method includes multi-scale training, testing and horizontal flip test.

YOLACT [36]: This method achieves good segmentation result by dividing the instance segmentation task into two parallel sub-tasks, one of which generates a set of masks, and the other predicts the mask coefficients of each instance.

Mask R-CNN [6]: This method applies ROIAlign pooling operation to achieve more accurate segmentation result.

Table 1. Comparison with some other algorithms on MS COCO dataset (%).

Methods	AP	AP ⁵⁰	AP ⁷⁰	AP ^S	AP ^M	AP ^L
MNC	24.6	39.9	19.4	4.7	25.9	43.6
FCIS + OHEM	29.2	49.5	-	7.1	31.3	50.0
FCIS + + OHEM	33.6	54.5	-	-	-	-
YOLACT	31.2	50.6	32.8	12.1	33.3	47.1
Mask R-CNN	35.7	58.0	37.8	15.5	38.1	52.4
RM-RCNN	37.2	55.5	40.9	18.1	43.7	56.5

Table 2. Comparison with some other algorithms on PASCAL VOC dataset (%).

Methods	AP	AP ⁵⁰	AP ⁷⁰	AP ^S	AP ^M	AP ^L
Fast R-CNN	20.5	44.3	24.8	4.1	20.0	35.8
Faster R-CNN	21.1	40.9	19.9	6.7	22.5	32.3
SIN	23.2	44.5	22.0	7.3	24.5	36.3
MIFNet	26.0	48.7	24.9	8.1	30.4	42.0
Mask R-CNN	38.2	60.3	41.7	20.1	41.1	50.2
RM-RCNN	40.5	62.7	43.3	23.5	42.8	51.5

Although aforementioned segmentation algorithms have achieved satisfactory results, they did not consider the interdependence between objects, and some pyramid-based network algorithms did not transmit messages between different layers. Therefore, it can be clearly seen from Table 1 that RM-RCNN achieves the AP of 37.2%, which is higher than other comparison methods. It is worth noting that AP50 and AP70 indicate that the threshold of IoU (Intersection over Union) is greater than 0.5 and 0.7, respectively. APS, APM and APL represent the AP of objects of different sizes (such as small, medium and large), respectively. By comparing the evaluation indexes of APS, APM and APL, we can clearly see that RM-RCNN achieves the optimal performance when dealing with the objects with different sizes. That's due to the pyramid network structure in our model can effectively capture the features with different scales. Mask R-CNN also adopts pyramid network as its backbone network, so its performance is better than other methods. Nevertheless, because RM-RCNN adds DCM in different layers of pyramid network and transmits semantic information of different layers through CSITM, the experimental results are superior to Mask R-CNN on MS COCO dataset.

Table 3. Comparison of different DCM combinations on different datasets. DCM_i stands for deploying DCM on the i -th layer.

The combinations of DCM	MS COCO (%)	PASCAL VOC (%)
[DCM_1 , DCM_2 , DCM_3 , DCM_4]	37.20	40.50
[DCM_1 , DCM_2 , DCM_3]	37.13	40.16
[DCM_1 , DCM_3 , DCM_4]	37.09	40.15
[DCM_2 , DCM_3 , DCM_4]	36.99	40.22
[DCM_1 , DCM_2]	36.80	40.10
[DCM_1 , DCM_3]	36.98	39.89
[DCM_1 , DCM_4]	36.72	39.45
[DCM_2 , DCM_3]	36.89	39.34
[DCM_2 , DCM_4]	36.02	39.02
[DCM_3 , DCM_4]	36.66	39.23
[DCM_1]	36.50	38.46
[DCM_2]	36.34	38.59
[DCM_3]	36.43	38.68
[DCM_4]	35.99	38.97

Table 4. Comparison of results of different CSITM combinations on different datasets. CSITM_i stands for deploying CSITM on the *i*-th layer.

The combinations of DCM	MS COCO (%)	PASCAL VOC (%)
[CSITM ₁ , CSITM ₂ , CSITM ₃]	37.20	40.50
[CSITM ₁ , CSITM ₂]	36.87	40.36
[CSITM ₂ , CSITM ₃]	36.36	40.04
[CSITM ₁ , CSITM ₃]	36.54	39.80
CSITM ₁	36.03	39.32
CSITM ₂	35.51	39.05
CSITM ₃	35.78	38.82

In Table 2, we compare the proposed RM-RCNN with some existing algorithms on PASCAL VOC dataset. The comparison methods are as follows:

Fast R-CNN [25]: This method uses deep convolutional neural network to classify objects and realize the regression of bounding boxes.

Faster R-CNN [26]: This method by introducing RPN greatly improves the generation of proposals and achieves the end-to-end object detection task.

SIN [18]: This method proposes a network of relationships between objects and good detection results are obtained.

MIFNet [19]: This method establishes the relationship between objects adaptively and the network is light and fast.

Although all of the above algorithms are impressive, and some algorithms also take the interdependence between objects into consideration, only single-scale features are modeled in them. Thus, they cannot effectively handle the information with different scales. For example, small scale features contain more semantic information, while large scale features contain more texture information. Thus, ignoring different information reflected by different scale features will limit their performance. From Table 2, it can be seen clearly that the AP obtained by our method is 40.5%, which is higher than other comparison methods. Moreover, our method obtains better results than other approaches under various IoU thresholds and object sizes. This is because that extracting features of different scales using pyramid network is beneficial to detection task, and making full use of features of different scales can extract more fine features.

In summary, through the comparative analysis with the classical segmentation and detection algorithms, it can be concluded that the RM-RCNN can extract more refined semantic information by using the intra-layer relationship calculation module and cross-scale information transmission module.

Ablation Studies. In order to justify the effectiveness of the proposed modules (DCM and CSITM) in our network, we perform several ablation studies.

Table 3 shows different combinations of DCM on the two datasets. We conduct 14 groups of experiments respectively. According to Table 3, it can be found that when DCM is added to a single layer, the AP obtained by the network becomes worse with the

increase of feature scale. This indicates that the semantic information obtained at the high level is richer, and the use of DCM at the high level can improve the network performance. However, among all combination strategies, our network obtains its optimal performance when deploying the DCMs on all layers.

Table 4 shows different combinations of CSITM on the two datasets. We conduct 7 groups of experiments respectively. According to Table 4, it can be found that the performance of our network can be improved by using CSITM. Furthermore, we can see that our method achieves its best performance when the CSITMs are deployed between all adjacent layers ([CSITM1, CSITM2, CSITM3]). This illustrates the effectiveness of information transmission in a pyramid structure.

Figure 6 shows some experimental results on MS COCO dataset. As can be seen from this figure, RM-RCNN can obtain accurate results for both semantic segmentation and object detection tasks.



Fig. 6. Some experimental results on MS COCO dataset.

5 Conclusion

Semantic segmentation and object detection are very important subjects in computer vision. In this paper, we propose a network based on Mask R-CNN for semantic segmentation and object detection tasks, which not only utilizes the rich features of inter-object dependence within the same layers, but also transmit information between different layers. By using the object dependency calculation module, the geometric and appearance features of objects can be employed to model the relationship information, so as to improve the accuracy of segmentation and detection. At the same time, the cross-scale information transfer module can refine the information between features of different levels, so that the network can effectively retain useful information and discard useless information. Experimental results on two benchmark datasets demonstrate the effectiveness of our method.

References

1. Goodfellow, I.J., Bengio, Y., Courville, A.C.: Deep learning. In: Adaptive Computation and Machine Learning. MIT Press (2016), <http://www.deeplearningbook.org>
2. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
3. Badrinarayanan, V., et al.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)

4. Chen, L.C., Papandreou, G., et al.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint [arXiv:1412.7062](https://arxiv.org/abs/1412.7062) (2014)
5. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. *Adv. Neural Inf. Proces. Syst.* **24**, 1–9 (2011)
6. He, K., Gkioxari, G., et al.: Mask R-CNN. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)
7. Mottaghi, R., et al.: The role of context for object detection and semantic segmentation in the wild. In: *Computer Vision and Pattern Recognition* (2014)
8. Yu, S., et al.: Democracy Does Matter: Comprehensive Feature Mining for Co-Salient Object Detection. arXiv preprint [arXiv:2203.05787](https://arxiv.org/abs/2203.05787) (2022)
9. Chen, Q., et al.: You only look one-level feature. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13039–13048 (2021)
10. Sun, P., et al.: Sparse R-CNN: end-to-end object detection with learnable proposals. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14454–14463 (2021)
11. Wang, J., et al.: End-to-end object detection with fully convolutional network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15849–15858 (2021)
12. Li, W., et al.: SIGMA: Semantic-complete Graph Matching for Domain Adaptive Object Detection. arXiv preprint [arXiv:2203.06398](https://arxiv.org/abs/2203.06398) (2022)
13. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: *IEEE International Conference on Computer Vision*, vol. 2, p. 273. IEEE Computer Society (2003)
14. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
15. Zeng, X., Ouyang, W., Yang, B., Yan, J., Wang, X.: Gated bi-directional CNN for object detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016. LNCS*, vol. 9911, pp. 354–369. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_22
16. Shrivastava, A., Gupta, A.: Contextual priming and feedback for faster R-CNN. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016. LNCS*, vol. 9905, pp. 330–348. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_20
17. Chen, X., Gupta, A.: Spatial memory for context reasoning in object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017)
18. Yong, L., Wang, R., Shan, S., Chen, X.: Structure inference net: Object detection using scene-level context and instance-level relationships. In: *IEEE* (2018)
19. Zhang, Y., Kong, J., Qi, M., Liu, Y., Lu, Y.: Object detection based on multiple information fusion net. *Appl. Sci.* **10**(1), 418 (2020)
20. Zhao, H., et al.: Pyramid scene parsing network. In: *IEEE Computer Society* (2016)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. LNCS*, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Pinheiro, P., Collobert, R., Dollar, P.: Learning to segments objects candidates. *Adv. Neural Inf. Proces. Syst.* **28**, 1–9 (2015)
23. Dai, J., He, K., Li, Y., Ren, S., Sun, J.: Instance-sensitive fully convolutional networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016. LNCS*, vol. 9910, pp. 534–549. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_32

24. 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)
25. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **28**, 1–9 (2015)
27. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. *Int. J. Comput. Vision* **104**(2), 154–171 (2013)
28. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. *Neural Process. Lett.* **9**(3), 293–300 (1999)
29. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016*. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
30. Redmon, J., et al.: 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)
31. Hochreiter, S., et al.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
32. Lin, T.-Y., et al.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
33. Everingham, M., et al.: The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision* **88**(2), 303–338 (2010)
34. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. *IEEE* (2016)
35. Li, Y., et al.: Fully convolutional instance-aware semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2359–2367 (2017)
36. Bolya, D., et al.: Real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157–9166 (2019)



Stitching High Resolution Notebook Keyboard Surface Based on Halcon Calibration

Gang Lv^{1,2,3}, Hao Zhao³, Zuchang Ma^{1,2}, Yining Sun^{1,2}, and Fudong Nian^{3(✉)}

¹ Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
lvgang@hfuu.edu.cn, {zcma, ynsun}@iim.ac.cn

² School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China

³ School of Advanced Manufacturing Engineering, Hefei University, Hefei 230601, China
nianfd@hfuu.edu.cn

Abstract. Aiming at the technical problems of poor stitching accuracy and difficulty in real-time stitching in the high resolution notebook keyboard surface image stitching algorithm, this paper proposes a high resolution notebook keyboard surface image stitching algorithm based on Halcon calibration. The method in this paper selects two cameras with different resolutions. First, the dot calibration board is used to calibrate the camera's internal parameters and 3D pose with Halcon; second, the 3D pose of the two cameras is mapped so that the two images are located in the same world coordinate system, and then the homography matrix is obtained; finally, the splicing fusion is performed. In addition, the method proposed in this paper can also be applied to image mosaic of more cameras. Extensive experiments demonstrate that compared with the widely-used SIFT, SURF and ORB-based methods, the proposed algorithm has several advantages: (1) compared with the algorithm with the smallest error, the error is reduced from 6.53 mm to 2.89 mm; (2) the homography matrix can be used directly, which greatly reduces the time of image stitching.

Keywords: Halcon · Image stitching · Coordinate system transformation · Homography matrix

1 Introduction

As an indispensable tool in today's life, notebooks are produced in large quantities every day. In the production process of notebooks, a series of inspections are required, and a large number of images are inevitably generated during the automatic inspection process. As part of the inspection work, the clarity and resolution of these images are very important. However, a single camera may not be able to capture the entire area of the notebook, and some areas require higher resolution and clarity, so in the production process, it is necessary to use multiple cameras to shoot and use image stitching technology to ensure real-time acquisition to meet inspection requirements notebook image. To obtain images that meet the requirements, the image stitching algorithm is extremely important, and it is required to meet the requirements of high precision and real-time performance, both of which are indispensable, otherwise, it will not meet the requirements of production inspection.

Image stitching algorithm [1–3] refers to the algorithm of stitching two or more images with overlapping areas into one image. The stitched image is required to have richer and more accurate scene information [4, 5], for example, it can obtain a wider field of vision or higher resolution. At present, image stitching is a popular algorithm based on feature matching, including SIFT [6, 7], SURF [9], and ORB [8, 10, 11]. However, the algorithm based on feature matching needs to extract the feature points of each group of images, which is very time-consuming, and there is bound to be the risk of false matching, resulting in the inability to ensure the splicing effect, and can not meet the real-time and high precision requirements in the automatic detection process of the notebook keyboard surface.

In addition, in the production process, the two cameras installed cannot be in the same horizontal plane, resulting in pictures taken by different cameras, not in the same 3D pose, that is, they cannot be in the same world coordinate system, which greatly reduces the accuracy of image splicing.

Aiming at the above difficult problems in image stitching, this paper proposes a high-resolution keyboard image mosaic algorithm based on Halcon, which can obtain the mosaic result image in real-time with guaranteed accuracy. Firstly, the dot calibration board is used to calibrate the camera [12, 13] to obtain the internal parameters and 3D pose of the camera. Secondly, the 3D pose is translated and rotated to eliminate the influence of the height of the calibration board itself, to obtain a new 3D pose, so that the two images can be in the same world coordinate system, which is transformed into an external parameter matrix, and then the homography matrix [14, 15] is calculated by using the internal parameter matrix and external parameter matrix. Finally, the homography matrix is used for perspective transformation to complete the image stitching. In addition to the stitching of two images, the algorithm proposed in this paper is also applicable to the stitching of multiple images at the same resolution. Taking the first camera as the benchmark, the 3D pose from other cameras to the first camera is calculated in turn, to ensure that all images are in the same world coordinate system, then obtain the homography matrix corresponding to each camera, complete image registration, and finally carry out image fusion to obtain the final stitching result.

2 Related Work

The core task of image mosaic is to get an appropriate homography matrix, then use the homography matrix for perspective transformation, and finally complete the image stitching. Homography matrix is a matrix used to describe the position mapping relationship between pixel coordinate system and world coordinate system, that is, the mapping matrix from one plane to another [16, 17]:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & T \\ \vec{0} & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} = AB \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} = H \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} \quad (1)$$

where H is the homography matrix, including the camera's internal parameter A and external parameter B , where the internal parameter belongs to the inherent attribute of

the camera. (u, v) is the coordinate of the midpoint of the pixel coordinate system, (u_0, v_0) represents the origin in the pixel coordinate system, and (X_w, Y_w, Z_w) is the point in the world coordinate system (assuming that the image is located in the plane of the world coordinate system $Z_w = 0$).

For the calibration of camera internal parameters, Zhang Zhengyou's calibration method [18, 19] is mostly used nowadays. Zhang Zhengyou's calibration method uses a checkerboard calibration board. Firstly, the camera image is binarized, and the candidate points of chessboard corners are found by finding quadrilateral (black chessboard area). Secondly, only those quadrilaterals that meet the specific size criteria [20] are filtered and organized in a regular grid structure, and the size of the grid structure matches the size specified by the user. Because the angle is infinitesimal mathematically, it is unbiased under perspective transformation or lens distortion. The checkerboard calibration board can determine the position of the corner with very high accuracy [21]. Using the checkerboard calibration board must ensure that the entire calibration board must be in the image, which also adds difficulty to the calibration process, because the edge information of the checkerboard is very important, and they properly constrain the distortion of the lens, even if a small part is missing, it will cause accuracy is reduced. However, the dot calibration plate [22] does not need to pay attention to these, and the high fitting degree of the center of the solid circle makes the calibration accuracy of the dot calibration plate higher.

In the study of image processing, circles can be detected as "spots" in an image. Some simple conditions, such as area, roundness, and convexity, can be applied to these binary speckle regions to filter out the bad feature points in the candidate points [23, 24]. After finding the appropriate candidate object, the pattern is recognized and filtered again by using the rule structure of features. In addition, because all pixels around the circle can be used to reduce the influence of image noise, the determination of the circle can be very accurate.

Most of the popular image stitching algorithms are based on feature extraction algorithms such as SIFT, SURF, and ORB. First, extract the feature points of the image to be spliced, then match the extracted feature points, use the matched feature points to calculate the homography matrix for image stitching, and finally complete the image fusion. However, in the process of feature point extraction [25, 26], there will inevitably be errors in feature point extraction and feature point matching errors, which will reduce the accuracy of image stitching. The method proposed in this paper does not need to perform feature point extraction and feature point matching, because accurate feature points and one-to-one correspondence are provided on the dot calibration board so that problems such as incorrect feature point extraction and feature point matching errors will not occur. At the same time, feature point extraction and matching take a lot of time, which brings great challenges to the real-time requirements of industrial production.

3 Our Approach

3.1 Camera Internal Parameter Calibration

Due to the influence of the camera itself and the lens, the image was taken by the camera often produces distortion. To remove the distortion, it is necessary to calibrate

the camera to obtain the internal parameters of the camera. Camera calibration is divided into three steps. The first step is to convert from the pixel coordinate system to the image coordinate system, as shown in formula (2); the second step is to convert from the image coordinate system to the camera coordinate system, as shown in formula (3); the third step is to convert the camera coordinate system to the world coordinate system, as shown in formula (4). It can be seen that finding the internal and external parameters of the camera is to find the projection transformation matrix [27] from the pixel coordinate system to the world coordinate system, as shown in formula (7):

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} dx & 0 & -u_0 dx \\ 0 & dy & -v_0 dy \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (2)$$

$$Z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (4)$$

where (x, y) represents the coordinates of the image coordinate system, (u, v) represents the coordinates of the pixel coordinate system, and (u_0, v_0) represents the origin in the pixel coordinate system. (X_c, Y_c, Z_c) is the point in the camera coordinate system. Formula (4) represents rigid body transformation, including rotation and translation, where (X_w, Y_w, Z_w) represents the coordinates of the midpoint of the world coordinate system, R_1, R_2, R_3 refers to the matrix obtained by rotating around the X, Y, and Z axes respectively, $R = R_1R_2R_3$ represents the rotation matrix, and T represents the translation matrix (Fig. 1).

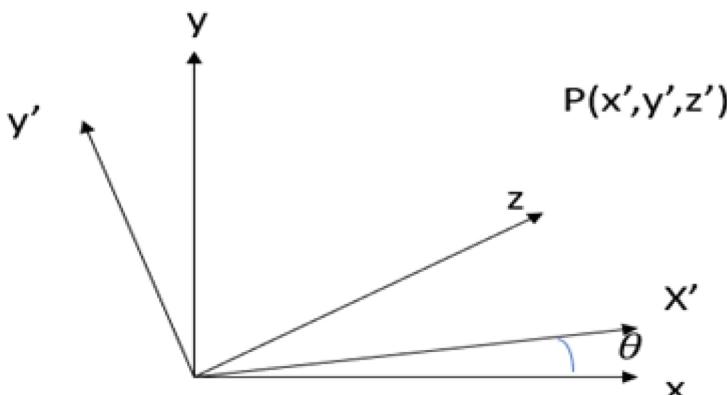


Fig. 1. Schematic diagram of rotating θ around the Z-axis.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R_1 \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \quad (5)$$

In the same way, rotate φ around the X-axis and select ω around the Y-axis to get R2, R3.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi \\ 0 & -\sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R_2 \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \cos\omega & 0 & -\sin\omega \\ 0 & 1 & 0 \\ \sin\omega & 0 & \cos\omega \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R_3 \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \quad (7)$$

The method in this paper uses the calibration tool that comes with Halcon and uses the dot calibration plate to automatically detect the dots in the image by collecting about 20 images with the calibration plate in different directions and positions and loading the calibration image in the calibration tool, and using the spatial position relationship satisfied by the calibration plane space dots and the corresponding image dots to establish the equation system as formula (2) (3) (4), and solve the camera internal parameter matrix distortion coefficient and 3D pose (Tables 1 and 2).

Table 1. Calibration results of internal parameters of camera 1

parameters	Calibration results
Internal parameter A1	[4640.9344315,0,2007.23;0, 4.640087,1494.68;0,0,1]
3D Pose	[-0.0124718, -0.0208803, 0.527218, 358.266, 357.826, 184.522, 0]
Distortion coefficient	[321.014, -2.10981e + 06, -2.16056e + 10, 0.0042086, -0.00672502]

Table 2. Calibration results of internal parameters of camera 2

parameters	Calibration results
Internal parameter A2	[7593.719743,0,1204.43;0, 7593.409091, 1016.74;0,0,1]
3D Pose	[-0.0069291, -0.0347552, 0.524949, 357.317, 11.5307, 359.921,0]
Distortion coefficient	[656.965,3.2321e + 06, 2.42197e + 11, 0.0428826, 0.0326885]

3.2 Transformation Mapping Between Coordinate Systems

Because image stitching can only be performed in the same coordinate system, this paper maps both images to the same world coordinate system [28]. The 3D pose obtained by calibration is obtained by mapping a group of two-dimensional coordinate points. The Z-axis defaults to 0, but the calibration board has a height, so it is necessary to map the 3D pose obtained from the calibration to eliminate the influence of the calibration board itself on the calibration. First, map the point in the upper left corner of image 1 taken by camera 1 to its corresponding world coordinate system, then translate the 3D pose PoseMatrix1 of camera 1, and translate the origin of the world coordinate system to the position of the point in the upper left corner of image 1 in the world coordinate system to obtain a new 3D pose NewOrigin_PoseMatrix1, such as formula (8):

$$\text{NewOrigin_PoseMatrix1} = \text{PoseMatrix1} \cdot \begin{bmatrix} 1 & 0 & 0 & \begin{pmatrix} Dx \\ Dy \\ Dz \end{pmatrix} \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

where Dx and Dy are the coordinates of the upper left corner pixel of image 1 in the world coordinate system, and Dz is the thickness of the calibration plate, in meters.

For camera 2, the origin of its world coordinate system is also translated to the origin of the world coordinate system where image 1 is located, so that the mapped image 2 and image 1 are located in the same world coordinate system, as shown in the formula (9):

$$\text{NewOrigin_PoseMatrix2} = \text{PoseMatrix2} \cdot \begin{bmatrix} 1 & 0 & 0 & \begin{pmatrix} Dx \\ Dy \\ Dz \end{pmatrix} \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (9)$$

3.3 Obtain Homography Matrix

The external parameter matrix required in this article mainly refers to the transformation matrix R from the world coordinate system to the camera coordinate system. The 3D pose PoseMatrix1 of camera 1 can be obtained through Halcon calibration, and then the new 3D pose NewOrigin_PoseMatrix1 is obtained through coordinate system transformation. Since each column vector of R represents the direction of each coordinate axis of the world coordinate system in the camera coordinate system; t represents the representation of the origin of the world coordinate system in the camera coordinate system. Therefore, the extrinsic parameter matrix of the camera can be obtained by inverting the 3D pose, as shown in formula (10):

$$\begin{bmatrix} R & t \end{bmatrix} = \begin{bmatrix} \text{NewOrigin_PoseMatrix1}^T & -\text{NewOrigin_PoseMatrix1}^T C \\ 0 & 1 \end{bmatrix} \quad (10)$$

Further formula (11) can be obtained:

$$R = \text{NewOrigin_PoseMatrix1}^T t = -RC \quad (11)$$

Similarly, the mapped external parameter matrix of camera 2 can be obtained $R2 = \text{NewOrigin_PoseMatrix}2^T$.

$$P1 = A1 * \text{NewOrigin_PoseMatrix}1^T \quad (12)$$

$$P2 = A2 * \text{NewOrigin_PoseMatrix}2^T \quad (13)$$

where P1 and P2 represent the final homography matrices of camera 1 and camera 2, respectively.

The above steps only need to be performed once on a production line, and the homography matrix obtained from the calculation is retained for repeated use.

3.4 Splicing Fusion

After obtaining the homography matrix of camera 1 and camera 2, map the images at different positions captured by camera 1 and camera 2 at the same time to obtain the mapping result map. After obtaining two mapping result maps, the mapping result map of camera 2 is processed in grayscale, and then the image is binarized by using Otsu threshold segmentation [29–31], to obtain the pixel value of the threshold area. Finally, it is replaced with the image taken by camera 1 to complete the image fusion.

3.5 Stitching of Multi-camera Images

The method proposed in this paper is also applicable to image stitching of multiple cameras of the same resolution. First, place multiple calibration boards on the same plane (each camera corresponds to one calibration board), obtain the internal parameters and 3D pose of each camera in the same way, and then select a reference camera to obtain the remaining cameras for the reference camera in turn. The mapping relationship of the world coordinate system, and finally complete the splicing.

$$H_2 = H_1 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, H_3 = H_1 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, H_{21} = H_2 \cdot H_3 \quad (14)$$

where H_1 is the 3D identity matrix, $t_1 = (x_1 + w, y_1 + h, dh)^T$, $t_2 = (-d_{21}, 0, 0)^T$, (x_1, y_1) is the origin of the reference camera in the world coordinate system, w is the width of the image taken by the reference camera in the world coordinate system, h is the width of the image taken by the reference camera in the world coordinate system, and dh is the width of the calibration board Thickness, d_{21} is the distance from the center of the second plate to the center of the benchmark calibration plate (the calibration plate corresponding to the benchmark camera), H_{21} is the mapping transformation matrix between the world coordinate system of the second camera and the world coordinate system of the reference camera. By analogy, the mapping transformation matrices of other cameras and the reference camera are obtained respectively, and then the final 3D pose is obtained, such as formula (15):

$$\text{New_Pose}2 = \text{Pose}2 \cdot H_{21} \quad (15)$$

where New_Pose is the new 3D pose, and Pose2 is the 3D pose obtained by the calibration of the second camera. Then, like the steps of dual cameras, the homography matrix is obtained by using a 3D pose for image mapping. Finally, after the mapping result of each image is obtained, the Otsu threshold segmentation method is used to complete the stitching of all images.



Fig. 2. The original image (left) and its mapping result (right) were obtained by camera 1.

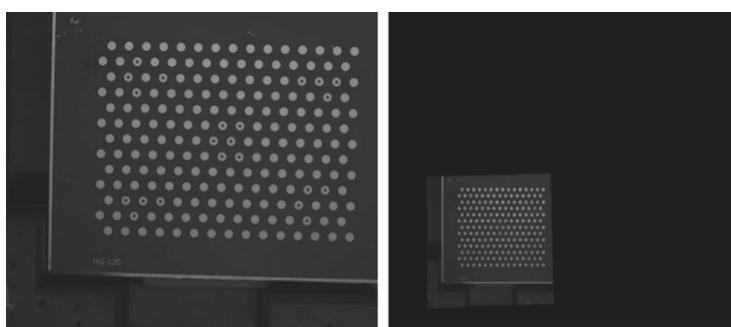


Fig. 3. The original image (left) and its mapping result (right) were obtained by camera 2.

4 Experiment

4.1 Experimental Details

The two cameras with different resolutions used in this article are both Hikvision industrial cameras. The model of camera 1 is HK-1614-10MP with a resolution of 4096 * 3000, and the model of camera 2 is MVL-HF0828M-6MP with a resolution of 4096 * 3000. 2448 * 2048. During the experiment, the calibration of the two cameras was completed first, and then the calibration board was placed at the label in the lower-left corner of the notebook, and the images containing the calibration board were captured by the two cameras at the same time, and the resolutions were 4096 * 3000 and 2448 * 2048 two images, image stitching, the results are shown in Figs. 2, 3, and 4.



Fig. 4. Schematic diagram of the result after image mosaic of the image of notebook keyboard surface with calibration board.

After completing the splicing of the notebook images with the calibration board, the two obtained homography matrices are saved and verified on the production line. This paper verifies the effect of the splicing algorithm on more than 200 different models of notebooks on the production line (covering notebooks of different thicknesses, notebooks of different sizes, and notebooks with different keyboard surfaces), as shown in Fig. 5. It can be seen that the stitched image obtained by the stitching algorithm proposed in this paper has high accuracy, and the stitching result has no ghosting, which meets the requirements of detecting the accuracy of image stitching on the notebook keyboard in the industrial production process.

4.2 Experimental Results

This paper verifies the image stitching methods based on SIFT, SURF, ORB, and the method of directly obtaining the homography matrix using the dots on the dot calibration board. The results are shown in Fig. 6 and Table 3.

Compare the splicing results of Figs. 5 and 6, and then compare them with Table 3, it can be concluded that the accuracy of the image mosaic method proposed in this paper is higher than the method of directly obtaining the homography matrix by using the dots on the dot calibration plate, and SIFT, SURF, and ORB-based methods.



Fig. 5. The result was obtained by using the splicing method in this paper

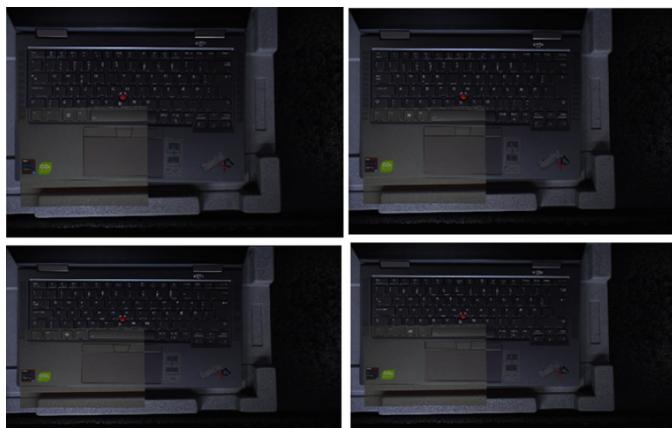


Fig. 6. The figure in the upper left corner is the mosaic result obtained by mapping the homography matrix directly obtained by Halcon using the dot calibration plate and the least square method; the other three figures respectively refer to the mosaic results obtained by extracting feature points using SIFT, SURF and ORB and matching the feature points.

The method proposed in this paper only needs to be calibrated once in a production line, and then the obtained homography matrix is directly applied to engineering. After a large number of experiments, the average processing time of image mosaic is 50ms after applying the homography matrix to the project. As shown in Table 3, the running time is far less than that required by the methods based on SIFT, SURF, and ORB. Due to the different focal lengths and placement of cameras on the production line, each different production line needs to be calibrated once, but the same algorithm can be used to obtain the homography matrix unique to the production line. Therefore, the splicing algorithm proposed in this paper has portability and repeatability.

Table 3. Algorithm performance comparison

Algorithm	Average error	Time
Method for obtaining homography matrix by directly using calibration plate	7.45 mm	0.0049 s
SIFT	6.53 mm	0.4861 s
SURF	7.12 mm	0.1323 s
ORB	9.33 mm	0.0142 s
Our	2.89 mm	0.0050 s

5 Conclusion

In summary, this paper proposes a high-precision keyboard surface image stitching algorithm based on Halcon calibration, which meets the high-precision and real-time requirements of automatic detection of notebook keyboard surfaces in industrial production. The method in this paper firstly uses Halcon for calibration to obtain the internal parameters and 3D pose, then performs the coordinate system mapping between cameras to obtain the final external parameter matrix, and finally uses the internal and external parameters to calculate the final homography matrix. In the method of this paper, the calibration of camera parameters is automated on Halcon, and the operation process is simple. One production line only needs to be calibrated once, which greatly reduces the complexity of camera calibration; the homography matrix can be reused, which greatly saves the code running time. In addition, the algorithm of this paper can be extended from the image stitching of two cameras to the image stitching of multiple cameras. Moreover, the algorithm in this paper is still applicable in many fields, such as unmanned driving, drone detection, and so on, and has practical industrial application prospects and significance.

Acknowledgments. This work was supported by the National Key R&D Program (No. 2020YFC2005603), and the Natural Science Research Project of Anhui Educational Committee (No. KJ2020A0651).

References

1. Xue, W., Zhang, Z., Chen, S.: Ghost elimination via multi-component collaboration for unmanned aerial vehicle remote sensing image stitching. *Remote Sens.* **13**(7), 1388 (2021)
2. Jung, K., Hong, J.: Quantitative assessment method of image stitching performance based on estimation of planar parallax. *IEEE Access*. **9**, 6152–6163 (2021)
3. Zhao, Q., Ma, Y., Zhu, C.: Image stitching via deep homography estimation. *Neurocomputing* **450**, 219–229 (2021)
4. Cao, W.: Applying image registration algorithm combined with CNN model to video image stitching. *J. Supercomput.* **77**(12), 13879–13896 (2021)

5. Wang, C., Gao, Z., Lu, Q.: Parallax-based color correction in image stitching. In: 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), pp. 69–74. IEEE (2020)
6. Rathi, K., Singh, P.: Copy move forgery detection by using integration of SLIC and SIFT. In: Jeena Jacob, I., Gonzalez-Longatt, F.M., Kolandapalayam Shanmugam, S., Izonin, I. (eds.) Expert Clouds and Applications. Lecture Notes in Networks and Systems, vol. 209, pp. 531–544. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2126-0_43
7. Hosseini-Nejad, Z., Agahi, H., Mahmoodzadeh, A.: Image matching based on the adaptive redundant keypoint elimination method in the SIFT algorithm. Pattern Anal. App. **24**(2), 669–683 (2021)
8. Bansal, M., Kumar, M., Kumar, M.: 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. Multimed. Tools App. **80**(12), 18839–18857 (2021)
9. Hasibuan, Z.A., Andono, P.N.: Contrast limited adaptive histogram equalization for underwater image matching optimization use SURF. J. Phys. Conf. Ser. **1803**(1), 012008 (2021)
10. Yang, K., Yin, D., Zhang, J.: An improved ORB algorithm of extracting features based on local region adaptive threshold. In: 6th International Conference on Systems and Informatics (ICSAI), pp. 1212–1217. IEEE (2019)
11. Yang, G., Chang, X., Jiang, Z.: A fast aerial images mosaic method based on ORB feature and homography matrix. In: 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–5. IEEE (2019)
12. Wang, G., Quan, W., Li, Y., Fang, S., Chen, H., Xi, N.: Fast and Accurate 3D Eye-to-hand calibration for large-scale scene based on HALCON. In: 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 230–234. IEEE (2021)
13. Yao, C., Yuan, Y., Li, J., Bi, L.: High precision tuning device of microwave cavity filter based on hand-eye coordination. In: 2019 Chinese Control Conference (CCC), pp. 7063–7068. IEEE (2019)
14. Tian, J., Wu, Y., Cai, Y.: A novel mosaic method for spaceborne ScanSAR images based on homography matrix compensation. Remote Sens. **13**, 2866 (2021)
15. Fan, R., Wang, H., Cai, P.: Learning collision-free space detection from stereo images: homography matrix brings better data augmentation. IEEE/ASME Trans. Mech. **27**, 225–233 (2021)
16. Ascencio, C.: Estimation of the Homography matrix to image stitching. In Applications of Hybrid Metaheuristic Algorithms for Image Processing, pp. 205–230 (2020)
17. Rashmi, C., Hemantha Kumar, G.: A parallel programming approach for estimation of depth in world coordinate system using single camera. In: Nagabhushan, P., Guru, D.S., Shekar, B.H., Kumar, Y.H.S. (eds.) Data Analytics and Learning. LNNS, vol. 43, pp. 77–91. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-2514-4_7
18. Wang, T., Wang, L.L., Zhang, W.G.: Design of infrared target system with Zhang Zhengyou calibration method. Opt. Precis. Eng. **27**(8), 1828–1835 (2019)
19. Wu, A., Xiao H., Zeng F.: A camera calibration method based on OpenCV. In: Proceedings of the 2019 4th International Conference on Intelligent Information Processing, pp. 320–324 (2019)
20. Wu, H., Wan, Y.: A highly accurate and robust deep checkerboard corner detector. Electron. Lett. **57**(8), 317–320 (2021)
21. Li, M., Liu, J., Yang, H.: Structured light 3D reconstruction system based on a stereo calibration plate. Symmetry **12**(5), 772 (2020)
22. Chuang, J.H., Ho, C.H., Umam, A.: Geometry based camera calibration using closed-form solution of principal line. IEEE Trans. Image Process. **30**, 2599–2610 (2021)

23. Yang, W.G., Qian, W., Qian, Y.: Camera internal parameter calibration based on rotating platform and image matching. In: Optics and Photonics for Information Processing XIII, vol. 11136, p. 111360Z (2019)
24. Bu, L., Huo, H., Liu, X.: Concentric circle grids for camera calibration with considering lens distortion. *Opt. Lasers Eng.* **140**, 106527 (2021)
25. Abdulhussain, S.H., et al.: A fast feature extraction algorithm for image and video processing. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
26. Nixon, M., Aguado, A.: Feature Extraction and Image Processing for Computer Vision. Academic Press, Cambridge (2019)
27. Simarro, G., Calvete, D., Plomaritis, T.A.: The influence of camera calibration on nearshore bathymetry estimation from UAV videos. *Remote Sens.* **13**(1), 150 (2021)
28. Nie, L., Lin, C., Liao, K.: Unsupervised deep image stitching: reconstructing stitched features to images. *IEEE Trans. Image Process.* **30**, 6184–6197 (2021)
29. Qingge, L., Zheng, R., Zhao, X.: An improved Otsu threshold segmentation algorithm. *Int. J. Comput. Sci. Eng.* **22**(1), 146–153 (2020)
30. El Khoukhi, H., Filali, Y., Yahyaouy, A., Sabri, M.A., Aarab, A.: A hardware implementation of OTSU thresholding method for skin cancer image segmentation. In: 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), pp. 1–5. IEEE (2019)
31. Tan, Z.Y., Basah, S.N., Yazid, H., Safar, M.J.: Performance analysis of Otsu thresholding for sign language segmentation. *Multimed. Tools App.* **80**(14), 21499–21520 (2021)



An Improved NAMLab Image Segmentation Algorithm Based on the Earth Moving Distance and the CIEDE2000 Color Difference Formula

Yunping Zheng¹(✉), Yuan Xu¹, Shengjie Qiu¹, Wenqiang Li¹, Guichuang Zhong¹, Mingyi Chen¹, and Mudar Sarem²

¹ School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510006, People's Republic of China
zhengyp@scut.edu.cn

² General Organization of Remote Sensing, Damascus, Syria

Abstract. How to effectively segment an image into the non-overlapping sub-regions and make the segmentation results conform to the perception of the human vision have always been a key issue in the field of computer vision. Extensive studies have proved that the CIEDE2000 formula is the most consistent method with the color space distribution for objects recognized by human perspective. The representation system of the NAMLab algorithm ignored the differences of the pixels at different positions of a region, and produced information loss in the process of dimensionality reduction. Therefore, in this paper, we propose an improved NAMLab algorithm based on the Earth Moving Distance (EMD) and the CIEDE2000 color difference formula. First, a K-means clustering is performed on the Lab features of all the pixels in the region in order to obtain the region color feature histogram. Then, the CIEDE2000 formula is used to calculate the differences between the cluster centers of the adjacent region color histograms. Finally, the EMD algorithm is used to calculate the color similarity among regions. When compared with the state-of-the-art algorithms, the experimental results presented in this paper demonstrate that the proposed algorithm has better segmentation performance, and the obtained segmentation results are more in line with the perception of the human vision.

Keywords: Hierarchical image segmentation · Image representation · Non-symmetry and Anti-packing pattern representation Model (NAM) · CIEDE2000 · Earth Moving Distance (EMD)

1 Introduction

Image segmentation refers to dividing an image into several disjoint regions according to features such as grayscale, color, spatial texture, geometric shape, etc., so that these features show consistency or similarity in the same region, while in different regions they show a significant difference. Image segmentation can be formulated as a classification problem of pixels with semantic labels (semantic segmentation) or partitioning of

individual objects (instance segmentation). Semantic segmentation performs pixel-level labeling with a set of object categories (e.g., human, car, tree, sky) for all image pixels, thus it is generally harder than an image classification, which predicts a single label for the entire image. Instance segmentation extends semantic segmentation scope further by detecting and delineating each object of interest in the image (e.g., partitioning of individual people) [1–4]. It is a crucial foundation of image processing technology and it is widely used in medical image analysis, remote sensing target extraction, fingerprint recognition, virtual reality, industrial automation and other fields. How to effectively segment an image into non-overlapping sub-regions and make the segmentation results conform to the perception of the human vision has always been a key issue in the field of computer vision [5–8].

Since J. Long et al. proposed the Fully Convolutional Neural network (FCN) [9], the image segmentation technology can be roughly divided into traditional methods and deep learning methods. The traditional image segmentation methods mainly include threshold methods [10], boundary detection methods [11], region methods [12], segmentation methods based on graph theory, segmentation methods based on clustering, etc. While the image segmentation methods based on deep learning mainly include: (1) Methods based on Convolutional Neural Network (CNN), including AlexNet [13], VGGNet [14], and ResNet [15]; (2) Methods based on Recurrent Neural Network (RNN), mainly including Long Short-Term Memory neural network (LSTM) [16]; (3) Image semantic segmentation methods based on region classification; (4) Image semantic segmentation methods based on pixel classification, where these methods are further divided into fully supervised learning methods and weak supervision learning methods, the former includes the FCN-based method DeepLab, the optimized convolution structure-based method Dilation10 [17], the encoder-decoder-based method Bayesian SegNet [18], and so on.

In recent years, image segmentation methods based on deep learning have been developed rapidly. For example, the FCN method adopted a cross-layer method, which not only took into account global semantic information and local position information, but also recovered the category of pixels from the abstract features. It was able to further extend the image level classification to the pixel level classification, and transform the network originally used for image classification into a network for image segmentation. The DeepLab method [19–21] added a fully connected conditional random field at the end of the FCN to optimize the boundary of the rough segmentation map, and used the Atrous convolution to expand the receptive field of the feature map, thus improving the solution of the problem that the FCN lacks spatial consistency and it is insensitive to the image details. However, compared with the image segmentation methods based on deep learning, the traditional image segmentation methods pay more attention to the color difference among the sub-regions and to the complementary information and hierarchical information between the texture and the shape features.

As for traditional image segmentation methods, Arbelaez et al. proposed an algorithm called gPb-OWT-UCM [22] for contour detection and image segmentation. In their method, the features were extracted by a multi-scale method on each channel, and they were converted into soft boundary maps of each position in each direction, and the possibility of each pixel as a boundary was calculated, that was, the weight of the pixel.

The fragmented boundary graph was then formed into an over-segmented superpixel map by the Oriented Watershed Transform (OWT), and each boundary was given a weight indicating the importance. Finally, the Ultra metric Contour Map (UCM) method was used to convert the regions into a hierarchical tree through greedy merging of regions to generate weighted contour images. Syu et al. proposed a new framework for hierarchical image segmentation based on iterative contraction and merging, called (ICM) [23]. Unlike the gPb-OWT-UCM algorithm, which extracted global information step by step through spectral clustering, the ICM used progressive region merging. As the iterative process progressed, the global information of the image was gradually explored, while focusing on the color, shape, and texture of the region merged during the contraction process. Compared with the gPb-OWT-UCM algorithm, the ICM algorithm was more efficient at maintaining better segmentation results. Recently, we proposed a hierarchical image segmentation algorithm based on the asymmetric inverse layout pattern representation model in the Lab color space called NAMLab [24]. The NAMLab defined the difference between two NAMLab-based regions, and iteratively performed NAMLab-based adjacent region merging algorithm in order to generate segmentation dendrogram step by step. As introduced in [25], the CIEDE2000 formula could better reflect the human vision's perception of chromatic aberration than the CIELab formula. However, considering that the original NAMLab algorithm took the average value of the Lab features of each pixel in the region, it ignored the difference of the pixels in different positions in the region, thus it had the disadvantage of information loss in the process of dimensionality reduction. Therefore, in this paper, we propose a regional color feature representation method based on K-means clustering algorithm and the Earth Move Distance (EMD) by using the CIEDE2000 formula [26] in order to replace the Euclidean distance and calculate the Lab color difference between different pixels. The improved NAMLab method proposed in this paper optimizes the measure of the chromatic aberration distance between the regions. And the segmentation results are more in line with the perception characteristics of the human vision.

The rest of this paper is organized as follows. In Sect. 2, we introduce the related concepts of the NAM and image features, as well as the improvements of the color similarity measure. In Sect. 3, we introduce the improved NAMLab algorithm and the modules based on the NAMLab framework. In Sect. 4, we present our experimental results illustrating the feasibility of the improvement. Finally, in Sect. 5, we summarize the ideas of this paper and discuss possible future work.

2 Related Concepts of NAM and Image Features

2.1 Non-symmetry and Anti-packing Pattern Representation Model

The Non-symmetry and Anti-packing pattern representation Model (NAM) [27–29] is an anti-packing problem. The idea of the NAM can be described as follows: Giving a packed pattern and n predefined sub-patterns with different shapes, pick up these n sub-patterns from the packed pattern, then represent the packed pattern with the combination of these sub-patterns.

The following are an abstract description of the NAM.

Suppose the original pattern is Γ , the reconstruction pattern is Γ' . Then, the NAM is a transform model from Γ to Γ' . The procedure of the transform can be written as follows:

$$\Gamma' = T(\Gamma) \quad (1)$$

where $T(\cdot)$ is a transform or encoding function.

The procedure of encoding can be obtained by the following expression:

$$\Gamma' = \cup_{j=1}^n p_j(v, A | A = \{a_1, a_2, \dots, a_{m_i}\}) + \varepsilon(d) \quad (2)$$

where Γ' is the reconstruction pattern; $P = p_1, p_2, \dots, p_n$ is a set of some predefined sub-patterns; n is the number of sub-pattern types; $p_j \in P$ is the j^{th} sub-pattern ($1 \leq j \leq n$); v is the value of p_j ; and $A = \{a_1, a_2, \dots, a_{m_i}\}$ is a parameter set of the sub-pattern p_j ($1 \leq j \leq n$). If the types of two sub-patterns are different, the numbers and the meanings of parameters in A are different.

2.2 Color Difference

A color space is a description of colors in a generally acceptable way under a certain standard, which is essentially an elaboration of coordinate systems and subspaces. Commonly used color spaces are RGB, HSV, CIELab, etc. The RGB is the most commonly used color space in image processing. However, since the RGB color space is sensitive to brightness, as long as the brightness changes, the components of the three channels will be changed accordingly. So, the RGB color space is no longer suitable for image processing. The CIELab color space can directly compare and analyze different colors by using the set distance in the color space, so it can be effectively and conveniently used to measure small color differences. Therefore, our proposed NAMLab algorithm uses the CIELab color space to calculate the color difference.

As for the difference between two pixels, the color difference formula used by the NAMLab algorithm is based on the Euclidean distance or the Gouraud distance.

1) The pixel-based Euclidean distance formula is as follows:

$$D_c(i, j) = \|C_i - C_j\|_2 \quad (3)$$

where $C_j = [L_i^*, a_i^*, b_i^*]^T$ and $C_j = [L_j^*, a_j^*, b_j^*]^T$ are the Lab color eigenvalues of the adjacent pixels i and j , respectively.

2) The pixel-based Gouraud distance formula is as follows:

$$D_c(i, j) = |g(x, y) - g_{est}(x, y)|_2 \quad (4)$$

Where $g(x, y)$ is the Lab color feature value of the current pixel, and $g_{est}(x, y)$ is calculated as follows:

$$g_{est}(x, y) = \begin{cases} g_5 + (g_6 - g_5)i_1, & x_1 < x_2, y_1 < y_2 \\ g_1 + (g_4 - g_1)i_2, & x_1 \neq x_2, y_1 = y_2 \\ g_1 + (g_4 - g_1)i_1, & x_1 = x_2, y_1 \neq y_2 \\ g_1, & x_1 = x_2, y_1 = y_2 \end{cases} \quad (5)$$

Among them, g_1, g_2, g_3 and g_4 are formed by the four corners of two pixels. $g_5 = g_1 + (g_2 - g_1)i_2, g_6 = g_3 + (g_4 - g_3)i_2, i_1 = (y - y_1)/(y_2 - y_1), i_2 = (x - x_1)/(x_2 - x_1)$.

As for the case between the two regions, the NAMLab algorithm uses the average value of the three components of the Lab feature of all the pixels in a region to represent the color feature of the region. Also, it uses the Euclidean formula to calculate the color distance between each pair of them.

1) The regional color features are represented as follows:

$$RC_i = \left[\frac{\sum_{j=1}^n L_j^*}{n}, \frac{\sum_{j=1}^n a_j^*}{n}, \frac{\sum_{j=1}^n b_j^*}{n} \right] \quad (6)$$

where RC_i is the Lab color feature of the region i , n is the number of the pixels in the region i , and L_j^*, a_j^*, b_j^* are the three component values of the Lab feature of the j^{th} pixel.

2) The region-based Euclidean formula is as follows:

$$D_c(i, j) = |RC_i - RC_j|_2 \quad (7)$$

where RC_i and RC_j are the Lab color eigenvalues of adjacent regions i and j , respectively.

Compared with the CIELAB, the CIEDE2000 [30] has made corrections in five aspects, including lightness weight function, chrome weight function, interactive term of chrome difference and chromatic aberration, and CIEDLAB's a^* factor. It further improved the color perception consistency of the color difference evaluation, and greatly improved the essence.

The CIEDE2000 color difference formula is as follows:

$$\Delta E = \sqrt{\left(\frac{\Delta L^*}{K_L S_L} \right)^2 + \left(\frac{\Delta C^*}{k_c S_c} \right)^2 + \left(\frac{\Delta H^*}{k_H S_H} \right)^2 + R_T f(\Delta C^* \Delta H^*)} \quad (8)$$

where ΔL^* , ΔC^* , and ΔH^* are the CIELAB metric lightness, chroma, and hue differences, respectively, which are calculated between the standard and the sample in each pair, $R_T f(\Delta C^* \Delta H^*)$ is an interactive term between the chroma and the hue differences, S_L , S_C , and S_H are the weighting functions for the lightness, chroma, and hue components, respectively, where the values calculated for these functions vary according to the positions of the sample pair being considered in the CIELAB color space, and k_L , k_C , and k_H values are the parametric factors to be adjusted according to different viewing parameters such as textures, backgrounds, separations, etc., for the lightness, chroma, and hue components, respectively.

2.3 Difference of Border, Texture, Size and Spatial Intertwining

In addition to color features, different image regions also have differences in boundary features, texture features, and size features. Therefore, in order to merge similar regions more reasonably, we need to consider these features.

Texture is a visual feature that reflects the homogeneous phenomenon in the image, which reflects the slowly changing or periodically changing the surface structure organization and the arrangement properties of the object surface. Texture is represented by the grayscale distribution of pixels and their surrounding spatial neighborhoods. J. Chen et al. used the Weber Local Descriptor (WLD) [31] to describe the texture features of image regions. The WLD is an efficient and robust texture descriptor, which is composed of a differential excitation operator ξ and an orientation operator θ .

The differential excitation ξ reflects the intensity information of the grayscale changes in the local window. It is obtained by calculating the ratio of the grayscale difference between the neighboring pixels and the central pixel in the local window to the grayscale value of the central pixel and using arctangent transformation as follows:

$$\xi(x_c) = \arctan\left[\frac{v_0}{v_1}\right] = \arctan\left[\sum_{i=0}^{p-1}\left(\frac{x_i - x_c}{x_c}\right)\right] \quad (9)$$

where x_c represents the intensity of the current pixel, x_i represents the intensity of the i^{th} adjacent pixel around the current pixel, and v_0 and v_1 represent the relative intensity difference and the current intensity of the current pixel and its neighbors, respectively.

The direction θ reflects the spatial distribution information of the grayscale change in the local window, which is described by the arctangent transformation of the grayscale difference between the adjacent pixels in the horizontal direction and the vertical direction of the local window, as follows:

$$\theta(x_c) = \arctan\left[\frac{v_2}{v_3}\right] \quad (10)$$

where v_2 and v_3 represent the intensity difference of the current pixel relative to the horizontal and the vertical directions, respectively. Then, the results of the differential excitation and direction are quantized into a $T \times D$ two-dimensional histogram, and converted into a $TD \times 1$ one-dimensional vector to form a texture feature vector W . The WLD feature difference $D_w(R_i, R_j)$ between the two regions R_i and R_j is described as follows:

$$D_w(R_i, R_j) = |W_{R_i} - W_{R_j}|_2 \quad (11)$$

At the same time, the color feature difference $D_{AB}(R_i, R_j)$ is considered, which is described as follows:

$$D_{AB}(R_i, R_j) = |AB_{R_i} - AB_{R_j}|_2 \quad (12)$$

The texture feature difference $D_T(R_i, R_j)$ between two regions is defined as:

$$D_T(R_i, R_j) = D_{AB}(R_i, R_j) * D_w(R_i, R_j) \quad (13)$$

As for the merging of regions, we have also considered the mean and the variance of the colors of the adjacent regions, and the boundary distance is described as follows:

$$D_B(R_i, R_j) = \frac{\sum_{p \in (R_i \cap B_{ij})} \sum_{q \in (R_j \cap B_{ij} \cap W_p)} |C_p - C_q|_2}{N_{pq}(R_i, R_j)} \quad (14)$$

where B_{ij} is the boundary region of R_i and R_j , p and q are the pixels on both sides of the boundary, and W_p is the local window around p .

In addition, an extra distance function is also considered, and it is defined according to the region size. The size distance is described as follows:

$$D_N(R_i, R_j) = \left(\frac{N_{R_i}^{1/t} N_{R_j}^{1/t}}{N_{R_i}^{1/t} + N_{R_j}^{1/t}} \right)^t \quad (15)$$

where N_R represents the total number of the pixels in the region R , and t , as the power of N_R , is an adjustable parameter.

In natural images, there may be some significant intensity or color variations on the same object, which may cause the same object to split multiplied by several intertwined regions in the spatial domain during pixel-based shrinking and merging. The spatial interleaving is described as follows:

$$D_I(R_i, R_j) = \min \left(\sum_{p \in R_i} f(MI_p, j), \sum_{q \in R_j} f(MI_q, i) \right) \quad (16)$$

where MI_p denotes the most common index in the local window at p pixels in the R_i region.

2.4 Region Difference

Based on the above statements and definitions, in order to comprehensively describe the characteristic information of the color, texture, size, boundary and spatial interweaving in the image region, the region distance formula is described as follows:

$$D(R_i, R_j) = D_N(R_i, R_j) \frac{(\alpha D_M(R_i, R_j) + \beta D_T(R_i, R_j) + \gamma D_u(R_i, R_j))}{\sqrt{\lambda + D_l(R_i, R_j)}} \quad (17)$$

where α , β and γ represent the adjustable weight parameters of the color, texture, and boundary information, respectively. The D_N terms are included to favor the merging of small regions into large ones, while the D_I terms are used to facilitate the merging of spatially intertwined regions.

3 An Improved NAMLab Algorithm

The improved NAMLab algorithm proposed in this paper includes four modules, namely, the representation module, the merging module, the removal module and the scanning module. In the representation module, the input image is divided into initial NAMLab blocks according to each pixel Lab feature based on the idea of the NAM. In the merging module, the eligible adjacent NAMLab blocks are further merged according to the Lab feature mean and variance. In the removal module, according to the texture, color, size and other characteristics of the NAMLab region, the eligible small regions and adjacent regions are iteratively merged. In the scanning module, the adjacent regions are merged according to the color distance based on the EMD distance and the equidistant texture size.

3.1 Representation Module

In the presentation module, the idea of the model is to represent the blocks of the input image by the mode of asymmetric inverse layout. An image is scanned line by line through raster scanning, and the distances between the adjacent pixels are judged according to the CIEDE2000 and the Gouraud formulas in order to expand the region, so that the original image is divided into one initial NAMLab rectangular block. Finally, the block map two-dimensional vector is used to record the NAMLab rectangular block number corresponding to each pixel, and its Lab feature mean and variance are also recorded.

3.2 Merging Module

In the merging module, for two adjacent NAMLab blocks, when the difference between the mean values of the two Lab features is less than a certain threshold μ and the variance of the merged regional Lab features is less than a certain threshold var, these two NAMLab blocks can be merged. The general process is as follows: Scan each NAMLab block in a raster way. For the current NAMLab block, first scan the NAMLab blocks of all adjacent pixels from the bottom to the top starting from the left side of the western border. If the NAMLab block to which the current adjacent pixel belongs is different from the current NAMLab block, then it is judged whether to merge the two NAMLab blocks according to the Union-Find algorithm. Then, scan all adjacent pixels from left to right starting from the northern boundary. If the NAMLab block to which the current adjacent pixel belongs is different from the current NAMLab block, it is judged whether to merge the two NAMLab blocks according to the Union-Find algorithm. This scanning is repeated until all adjacent pixels starting from the northern boundary have been processed.

3.3 Removal Module and Scanning Module

During the merging process of the NAMLab blocks, there are some small residual regions whose color mean and variance are quite different from the color mean and variance of

their adjacent rejoins so that they cannot be merged. Therefore, we customize a threshold for the size of the region. When the size of the current region is smaller than the threshold, the current region will be merged into a region with the least difference among all the neighboring regions.

In the scanning module, for the color features, we firstly cluster the Lab color feature of each NAMLab block to generate a Lab color feature histogram, then we calculate the distance between the histograms of each pair of the adjacent NAMLab blocks which is called the color difference through the EMD and the CIEDE2000 formulas. As for the texture features, we compute the texture distance between two NAMLab blocks using a WLD descriptor-based method. Then, we iteratively shrink and merge the remaining NAMLab blocks by combining the sizes of different NAMLab blocks with the spatial interleaving features.

3.4 Complexity Analysis

Compared with the original NAMLab, the improved NAMLab+ algorithm has the same complexity. In the first stage, the NAMLab+ only need to visit each pixel of the image one time, therefore the time complexity is $O(N)$ where N is the number of image pixels. In the second stage, assuming that K represents the number of regions after using the removal module, where K is much smaller than N , in the scanning module, we use a neighbor table to record all the information of all neighboring regions. Each time a pair of regions is merged with the smallest difference, the cost will be $O(K)$. Then, we repeat the fusion of the remaining $K-1$ regions until there is only one region to exit. During the fusion of the regions, the Union-Find algorithm is used for encoding. In the case of K image regions, the cost of the Union-Find algorithm is $O(K \alpha(K))$. When judging the region distance, the EMD algorithm first performs K-means clustering on all pixels in the region. Since the average number of elements in each cluster center is C , the Lab feature has three channels, and the number of iterations is I , so the cost is $O(60KI)$. Therefore, the total complexity of the NAMLab+ is $O(N + 60KI)$ which is equal to $O(N)$.

4 Experimental Results and Discussions

4.1 Evaluation Metrics and Datasets

In order to confirm that our proposed changes could improve the segmentation quality of the NAMLab algorithm for color images, we have conducted some experiments to compare the improved NAMLab algorithm, which we call it as the NAMLab+ algorithm, with the original NAMLab algorithm. Also, in our experiments, we have compared NAMLab+ with some popular image segmentation algorithms such as the ICM and the gPb-OWT-UCM algorithms.

We have used six popular public datasets for image segmentation named BSDS300, BSDS500, MSRCv2 [32], VOC2012 [33], SBD [34] and NYUv2 [35], which are used in the extensive published literatures.

4.2 Performance Evaluation

According to the popular evaluation indicators, i. e., CR, PRI, and VI, we have used the ODS and OIS criteria in order to select the segmentation layers for image segmentation, where the ODS is to find the optimal layer of the segmentation layers based on all images in the dataset and the OIS is to find the optimal segmentation layer for each image. By adjusting the parameters whose values are based on experience used in region distance formula of the proposed NAMLab+, where α is set to 1.18 increased by 0.18 compared with the NAMLab algorithm, we have obtained better segmentation results, indicating that our improvement has made the color feature of the image more effective in the segmentation process. The experimental results in Table 1 show that the image segmentation quality of the ICM algorithm is slightly worse than that of the gPb-OWT-UCM algorithm; while the results of the NAMLab algorithm and the ICM algorithm are roughly similar. Hence, we can see that the improved NAMLab+ algorithm has better performance than the NAMLab algorithm in general.

Table 1. Benchmarks on BSDS300 and BSDS500 datasets.

Algorithm	BSDS300					
	CR		PRI		VI	
	ODS	OIS	ODS	OIS	ODS	OIS
gPb-OWT-UCM	0.588	0.646	0.808	0.852	1.653	1.466
ICM	0.563	0.65	0.789	0.854	1.779	1.455
NAMLab	0.565	0.641	0.789	0.849	1.796	1.472
NAMLab+	0.564	0.647	0.788	0.851	1.776	1.474
Algorithm	BSDS500					
	CR		PRI		VI	
	ODS	OIS	ODS	OIS	ODS	OIS
gPb-OWT-UCM	0.588	0.647	0.827	0.856	1.69	1.475
ICM	0.571	0.648	0.814	0.857	1.762	1.472
NAMLab	0.571	0.642	0.805	0.854	1.817	1.491
NAMLab+	0.566	0.643	0.807	0.856	1.779	1.49

In addition, we have also tested the image segmentation performance of the gPb-OWT-UCM algorithm, the ICM algorithm, the NAMLab algorithm and the NAMLab+ algorithm on other four public datasets. The experimental results in Table 2 show that the performance of the improved NAMLab+ algorithm is very competitive to the NAMLab algorithm and the ICM algorithm.

Although the gPb-OWT-UCM algorithm has the highest image segmentation quality in general, its running time and memory consuming are much larger than those of the ICM algorithm, the NAMLab algorithm, and the NAMLab+ algorithm.

Table 2. Region benchmarks on other datasets.

Algorithm	CR		PRI		VI		Datasets
	ODS	OIS	ODS	OIS	ODS	OIS	
gPb-OWT-UCM	0.65	0.742	0.779	0.845	1.273	0.981	MSRCv2
ICM	0.642	0.748	0.748	0.854	1.205	0.958	
NAMLab	0.65	0.749	0.752	0.852	1.198	0.955	
NAMLab+	0.649	0.74	0.778	0.844	1.28	0.991	
gPb-OWT-UCM	0.582	0.641	0.862	0.892	1.877	1.617	SBD
ICM	0.581	0.66	0.855	0.9	1.859	1.561	
NAMLab	0.571	0.653	0.852	0.897	1.885	1.578	
NAMLab+	0.572	0.652	0.851	0.896	1.894	1.587	
gPb-OWT-UCM	0.65	0.713	0.654	0.752	0.972	0.922	VOC2012
ICM	0.67	0.737	0.669	0.78	0.835	0.79	
NAMLab	0.67	0.74	0.669	0.784	0.835	0.786	
NAMLab+	0.67	0.74	0.669	0.783	0.835	0.786	
gPb-OWT-UCM	Out of memory						NYUv2
ICM	0.453	0.502	0.845	0.862	2.592	2.346	
NAMLab	0.447	0.495	0.842	0.858	2.573	2.325	
NAMLab+	0.446	0.496	0.841	0.858	2.577	2.322	

Figure 1 shows the partial hierarchical image segmentation results of the ICM algorithm, the NAMLab algorithm and the improved NAMLab+ algorithm on the BSDS500 dataset. Figure 1 (a) is the original images. Figure 1 (b), (c), and (d) are the results of ICM algorithm, NAMLab algorithm, and NAMLab+ algorithm based on ODS over CR, respectively. Figure 1 (e), (f), and (g) are the results of ICM algorithm, NAMLab algorithm, and NAMLab+ algorithm based on OIS over CR, respectively. It can be seen that both the ICM algorithm and the NAMLab algorithm have good image segmentation performance. However, the region boundaries generated by the ICM algorithm sometimes notice so many details that the objects belonging to the same category are unnecessarily divided into two categories whereas the NAMLab algorithm sometimes confuses objects of two different classes. Compared with them, our improved NAMLab+ algorithm can accurately segment different regions. The segmentation results are closer to the perception of human vision, and have good consistency in the distribution of similar colors.



Fig. 1. Visual comparison of segmentation results generated by the ICM algorithm, the NAMLab algorithm, and the NAMLab+ algorithm on the BSDS500 dataset.

As stated above, the experimental results in this section verify the results of the theoretical analyses.

5 Conclusion and Future Work

In this paper, we have proposed an improved NAMLab algorithm based on the EMD Distance and the CIEDE2000 color difference formula. First, the K-means clustering is performed on the Lab features of all pixels in the region in order to obtain the region color feature histogram. Then, the CIEDE2000 formula is used to calculate the difference between the cluster centers of the adjacent region color histograms. Finally, the EMD algorithm is used to calculate the color similarity between the regions. Compared with the state-of-the-art algorithms, the experimental results presented in this paper demonstrate that the proposed algorithm has better segmentation performance, and the obtained segmentation results are more in line with the perception of the human vision. Although our algorithm achieves good segmentation results, there is still much room

for improvements. Whether it is to change the sub-pattern based on the NAM idea to transform the representation structure of the image or to modify the fusion method of different NAMLab blocks, our algorithm can be further improved to a certain extent. In the future, we will combine different region fusion methods for better rendering. We also believe that our algorithm can be further improved for better segmentation quality and efficiency.

Acknowledgment. This work is supported by the Natural Science Foundation of Guangdong Province of China under Grant No. 2017A030313349 and No. 2021A1515011517, the National Natural Science Foundation of China under Grant No. 61300134, and the National Undergraduate Innovative and Entrepreneurial Training Program under Grant No. 202110561070 and No. 202110561066.

References

1. Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D.: image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3523–3542 (2022). <https://doi.org/10.1109/TPAMI.2021.3059968>
2. Pandey, R., Lalchhanhima, R.: Segmentation techniques for complex image: review. *Int. Conf. Comput. Perform. Eval. (ComPE)* **2020**, 804–808 (2020)
3. Singh, V., Gupta, S., Saini, S.: A methodological survey of image segmentation using soft computing techniques. *Int. Conf. Adv. Comput. Eng. App.* **2015**, 419–422 (2015)
4. Sevak, J.S., Kapadia, A.D., Chavda, J.B., Shah, A., Rahevar, M.: Survey on semantic image segmentation techniques. *Int. Conf. Intell. Sustain. Syst. (ICISS)* **2017**, 306–313 (2017)
5. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging.* **39**, 1856–1867 (2020)
6. Hu, Q., et al.: RandLA-Net: efficient semantic segmentation of large-scale point clouds. *IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)* **2020**, 11105–11114 (2020)
7. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1483–1498 (2021)
8. Fan, D., et al.: Inf-Net: automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans. Med. Imaging.* **39**, 2626–2637 (2020)
9. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 640–651 (2017)
10. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979)
11. Davis, L.S.: A survey of edge detection techniques. *Comput. Graph. Image Process.* **4**, 248–270 (1975)
12. Adams, R., Bischof, L.: Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 641–647 (1994)
13. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. *Neural Inf. Process. Syst.* **25**, 1–9 (2012). <https://doi.org/10.1145/3065386>
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint: [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *IEEE Conf. Comput. Vision Pattern Recog. (CVPR)* **2016**, 770–778 (2016)

16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
17. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122* (2015)
18. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2015)
19. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs (2014). arXiv preprint [arXiv:1412.7062](https://arxiv.org/abs/1412.7062)
20. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018)
21. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous convolution for semantic image segmentation (2014). arXiv preprint: [arXiv:1412.7062](https://arxiv.org/abs/1412.7062)
22. Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 898–916 (2011)
23. Syu, J., Wang, S., Wang, L.: Hierarchical image segmentation based on iterative contraction and merging. *IEEE Trans. Image Process.* **26**, 2246–2260 (2017)
24. Zheng, Y., Yang, B., Sarem, M.: Hierarchical image segmentation based on nonsymmetry and anti-packing pattern representation model. *IEEE Trans. Image Process.* **30**, 2408–2421 (2021)
25. Gomez-Polo, C., Munoz, M.P., Luengo, M.C.L., Vicente, P., Galindo, P., Casado, A.: Comparison of the CIElab and CIEDE2000 color difference formulas. *J. Prosthet. Dent.* **115**, 65–70 (2016)
26. Luo, M.R., Cui, G., Rigg, B.: The development of the CIE 2000 colour-difference formula: CIEDE 2000. *Color. Res. Appl.* **26**, 340–350 (2001)
27. Zheng, Y., Yu, Z., You, J., Sarem, M.: A novel gray image representation using overlapping rectangular NAM and extended shading approach. *J. Visual Commun. Image Represent.* **23**, 972–983 (2012)
28. Liang, H., Zhao, S., Chen, C., Sarem, M.: The NAMlet transform: a novel image sparse representation method based on non-symmetry and anti-packing model. *Signal Process.* **137**, 251–263 (2017)
29. Zheng, Y., Sarem, M.: A fast region segmentation algorithm on compressed gray images using non-symmetry and anti-packing model and extended shading representation. *J. Vis. Commun. Image Represent.* **34**, 153–166 (2016)
30. Chou, C., Liu, K.: Perceptually optimized color image watermarking scheme based on CIEDE2000 color difference equation. In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 526–529 (2004)
31. Jie, C., Shan, S., Chu, H.: WLD: a robust local image descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1705–1720 (2010)
32. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: Proceedings of the British Machine Vision Conference, U.K., University of Warwick, pp. 55.1–55.10 (2007)
33. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010)
34. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1–8 (2009)
35. Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.): Computer Vision – ECCV 2012. LNCS, Springer, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-33709-3>



An Improved NAMLab Algorithm Based on CIECDE2000 Color Difference Formula and Gabor Filter for Image Segmentation

Yunping Zheng^(✉), Shengjie Qiu, Jiehao Huang, Yuan Xu, Zirui Zou,
and Pengcheng Sun

School of Computer Science and Engineering, South China University of Technology,
Guangzhou 510006, China
zhengyp@scut.edu.cn

Abstract. Image representation is an important problem in the field of computer vision. The Non-symmetry and Anti-packing pattern representation Model (NAM) is an effective pattern representation model. In order to further improve the image segmentation quality and segmentation efficiency, in this paper, we improve the recently published NAMLab algorithm in two aspects. First, the CIEDE2000 color difference formula is used to replace the calculation formula of the color feature similarity in the original NAMLab algorithm. The formula is based on the human vision response to RGB and it is used to accurately represent the reception of the color. Secondly, the calculation formula of texture features in the original NAMLab algorithm is modified. The original NAMLab algorithm is based on the Weber Local Descriptor (WLD) texture descriptor to describe the feature texture of the image. In order to better meet the characteristics of human vision observation, we found that the Gabor wavelet is very similar to the stimulus response of simple cells in the human visual system, which is more in line with the characteristics of the human vision, so we choose Gabor filter as the feature texture description of the image. Finally, the improved algorithm is compared with the state-of-the-art algorithms in the field of image segmentation on six datasets, and it achieves better results in terms of visual presentation and the segmentation indicators.

Keywords: NAMLab · CIEDE2000 · Gabor Filter

1 Introduction

Image segmentation refers to the process in which a computer divides an image into different blocks according to the relative consistency of the characteristics of different regions of the image. It is one of the basic problems in the field of computer vision and image processing. The preprocessing problems involved in specific applications such as scene understanding, image compression, and augmented reality are widely used in many fields such as medical images, traffic road recognition, underwater exploration, satellite remote sensing, and face recognition. A good image representation method can not only effectively compress the image and save the storage space, but also can improve the

speed of subsequent image processing. The image segmentation methods can be mainly divided into traditional image segmentation methods and image segmentation methods based on deep learning. With the gradual deepening of deep learning technology in recent years, the development of image segmentation methods based on deep learning is in full swing. While the performance is improved and the evaluation results are excellent, it also represents a new generation of the image segmentation model. At present, many great breakthroughs have been made in the field of image semantic segmentation based on deep learning. However, a large number of pixel-level annotations usually consume a lot of time, money and manpower. Therefore, insufficient or missing training data has become one of the key factors restricting the further development of image semantic segmentation.

The basic multi-level threshold technology has a large computational complexity and there is still room for improvement in segmentation accuracy. The Whale Optimization Algorithm (WOA) effectively balanced the exploration and the development process of the algorithm [1]. The traditional multi-phase image segmentation algorithm has long operation time, high computational cost and high error rate. Guo et al. proposed a model combining a multiphase image segmentation enhancement algorithm and a clustering-based algorithm [2]. This method reduced the sensitivity of the clustering algorithm during initialization, which made it easier for the multiphase image segmentation model under the clustering algorithm to segment the ideal image. At the same time, the image segmentation model could quickly obtain the minimum value, reduced the amount of calculation, and fully and effectively improved the efficiency. Thresholding methods for one-dimensional and two-dimensional had the disadvantage of high computational complexity, while Ashish et al. proposed a context-based three-dimensional Otsu algorithm, which was based on the mean error (ME) [3]. The performance indicators such as mean square error (MSE), peak signal-to-noise ratio (PSNR), feature similarity index (FSIM), structural similarity index (SSIM) and entropy have advantages over histogram-based Otsu. Monemian and Rabbani proposed a novel segmentation algorithm for optical coherence tomography images based on pixels intensity correlations [4], which used the intensity information of the pixels to find the distinguishing features of the boundary pixels located on the boundary of the retinal layer. Milano et al. proposed a definition of homogeneity and its quantification algorithm [5]. In recent studies, some segmentation algorithms such as the threshold-based watershed segmentation method [6], the edge-based image information detection [7], the threshold-based and clustering-based segmentation [8], the image processing technology based on edge detection and graph segmentation [9], the region-based image segmentation [10], were also being investigated.

In the case of the traditional image segmentation algorithms, the gPb-OWT-UCM algorithm [11] and the iterative contraction and merging (ICM) algorithm [12] were more representative. However, the gPb-OWT-UCM algorithm has some efficiency disadvantages due to its high time and space complexity. The ICM algorithm is a new framework

for hierarchical image segmentation. Compared with the gPb-OWT-UCM, the implementation of the ICM algorithm has more efficiency advantages. In 2021, Zheng et al. proposed the Nonsymmetry and Antipacking pattern representation Model (NAM) in the Lab color space for the hierarchical image segmentation [13], which is short for the NAMLab algorithm. In this paper, to further improve the image segmentation quality and segmentation efficiency, we proposed an improved version of the NAMLab algorithm, which is called as the NAMLab+ algorithm. The innovations are mainly reflected in the following two aspects.

First, on the basis of the NAMLab algorithm, the color difference formula is replaced. The original Euclidean distance formula is replaced with the CIEDE2000 color difference formula [14] since it can perform more precise calculation of color similarity. According to the research results in reference [15], the CIEDE200 chromatic aberration formula can better reflect the human vision's perception of the chromatic aberration differences.

Second, on the basis of the NAMLab algorithm, the texture formula is replaced. The original Weber Local Descriptor (WLD) is replaced with a Gabor filter to measure the texture's difference between regions since the Gabor filter has good time-frequency localization characteristics, and it is similar to the two-dimensional receptive field model of simple cells in the visual epidermis of most mammals (including humans).

The rest of this paper is organized as follows: Sect. 2 introduces the related concepts of the NAM and some proposed improvements. Section 3 introduces the architecture of the NAMLab+ algorithm. Section 4 gives the comparison results of the experimental results of the NAMLab+ algorithm and the state-of-art image segmentation algorithms on some public and popular datasets. Section 5 draws a conclusion.

2 Related Concepts and Proposed Improvements

2.1 Description of NAM

Suppose an original image pattern is Γ , two reconstructed non-distortion and distortion image patterns are Γ' and Γ'' , respectively. Then, the NAM is either a non-distortion model from Γ to Γ' or a distortion one from Γ to Γ'' . The procedure of the transform can be written by Eq. (1).

$$\Gamma' = T(\Gamma), \quad \Gamma'' = T(\Gamma) \quad (1)$$

where $T()$ is a transform or encoding function.

The procedure of the non-distortion encoding can be obtained by the following expression.

$$\Gamma' = \bigcup_{j=1}^n p_j(v, A | A = \{a_1, a_2, \dots, a_{m_i}\}) + \varepsilon(d) \quad (2)$$

where Γ' is the reconstructed image pattern; $P = \{p_1, p_2, \dots, p_n\}$ is a set of some predefined subpatterns; n is the type number of the subpatterns; $p_j (1 \leq j \leq n)$ is the j^{th} subpattern; v is the value of p_j ; A is a parameter set of p_j ; $a_i (1 \leq i \leq m_i)$ is a parameter set of shapes of p_j ; m is the serial number of p_j ; $\varepsilon(d)$ is a residue image pattern, and d is a threshold of $\varepsilon(d)$.

If the residue image pattern $\varepsilon(d)$ is removed from the non-distortion image pattern, then the distortion image pattern can be obtained by Eq. (3).

$$\Gamma'' = \bigcup_{j=1}^n p_j(v, A | A = \{a_1, a_2, \dots, a_{m_i}\}) \quad (3)$$

2.2 Color Difference

The color difference is an indicator introduced to measure the similarity of colors between two regions. In different color spaces, there are different ways to measure the color difference. Commonly used color spaces include the RGB color space, the CIELab color space and the CMYK color space, etc. The RGB color space is mainly used in the display system. The CMYK color space also has such device color-dependent characteristics. Not only any color in nature can be expressed through the CIELab color space, but also the way to express the color is to use the digitals. It has device-independent characteristics, and the color information that can be described by the other two color spaces can be mapped to the CIELab color space.

The calculation of the color difference in the original NAMLab algorithm only uses the Euclidean distance and the Gouraud distance [19] to calculate the color difference between the two regions. Therefore, not only the image representation and processing effect in the CIELab color space is better, but also a measure of color differences between regions is more refined.

Therefore, in this paper, our improved NAMLab+ algorithm will use the CIEDE2000 color difference formula based on the CIELAB color space to describe and calculate more refined color differences.

2.3 Gabor Filter

In addition to the similarity of color features, there are also differences in texture features and boundary features between image regions. As a feature that reflects the distribution structure of the image, texture plays a key role in the representation of the image. The selection of texture formula directly affects the human vision's observation of the image. The texture formula of the original NAMLab algorithm uses the WLD, which consists of a differential excitation operator and a direction operator. The differential excitation

operator calculates the ratio between the gray value of the domain pixel and the center pixel in the local window. The ratio generated by the arctangent transformation is designed to reflect the intensity information of the grayscale change in the local window. The direction operator is generated by the ratio of the grayscale values of the domain pixels in the horizontal direction and the vertical direction in the local window, and the arctangent transformation. It aims to reflect the spatial information of grayscale changes within a local window.

Therefore, in this paper, our improved NAMLab+ algorithm will use Gabor filter [20, 21] as the extraction of image texture features. In image processing, the Gabor function is used as the filter for feature extraction, and its expression in frequency and direction is related to the stimulus response of simple cells in the human visual system. Gabor filter is also very suitable for the expression and separation of image texture. The Gabor filter can be defined as a sine wave multiplied by a Gaussian function, and its Fourier transform is the convolution of the Fourier transform of its harmonic function and the Fourier transform of the Gaussian function. The complex expression of the Gabor filter is Eq. (4).

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right) \quad (4)$$

The real numbers are expressed by Eq. (5).

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \quad (5)$$

Imaginary numbers are expressed by Eq. (6).

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \sin\left(2\pi \frac{x'}{\lambda} + \psi\right) \quad (6)$$

where $x' = x \cos \theta + y \sin \theta$, $y' = -x \sin \theta + y \cos \theta$. The symbol λ measures the size of the extracted features. The symbol θ specifies the direction of the Gabor function fringes. The phase offset ψ is generally between -180° and 180° . The symbol γ is the spatial aspect ratio. The symbol b represents the bandwidth. In this paper, a Gabor filter with 6 scales and 8 directions is used to generate 48 feature maps of the same size, and each pixel of the image corresponds to a 48-dimensional Gabor feature vector.

The pseudo code of the Gabor filter algorithm in this paper is as follows:

Algorithm Gabor filter

Input: *Image*

- 1: Get x, y from image
- 2: Let $x' \leftarrow x\cos\theta + y\sin\theta$ and $y' \leftarrow -x\sin\theta + y\cos\theta$
/* x : The abscissa of the pixel.
/* y : The ordinate coordinate of the pixel.
- 3: Set scales₆ and direction₈ $\leftarrow [0, 2\pi, \pi/4]$
/* Stores six scales and eight directions for recall.
- 4: Set G_{68} and *geometric* $\leftarrow 1$
/* G store 48 – dimensional eigenvectors, *geometric* store

their product. */

- 5: Set i and $j \leftarrow 0$
/* Iterate through the six scales and eight directions of gabor.
*/
- 6: **for** $i \in [1, 8]$ **do**
/* theta: **Gabor** direction eight in total. */
- 7: **foreach** $j \in \text{scales} - \text{scales}_6$ **do**
 /* Gabor **scales** six in total. */
- 8: $G_{68} \leftarrow$ using Gabor filter
 /* The **Gabor** eigenvector is obtained using the Gabor
 filter formula. */
- 9: *geometric* $\leftarrow i\text{geometric} * G_{68}$
 /* Get **Gabor** eigenvectors of 48 dimensions. */
- 10: **end**
end
- 11: Let $g \leftarrow \sqrt[48]{\text{geometric}}$
/* Geometric means are used to average 48 – dimensional
Gabor feature vectors. */
- 12: Set four **coordinate** x_1, x_2, y_1, y_2 ;
Set six vector $g_1, g_2, g_3, g_4, g_5, g_6$;
/* Based on the Gouraud distance, the color characteristic value
of the current region is **calculated**, the following are the

calculation formulas. */

- 13: **if** $x_1 < x_2$ and $y_1 < y_2$ **then**
- 14: Let $g_{est} \leftarrow g_5 + (g_6 - g_5)y - y_1/y_2 - y_1$
end
- 15: **if** $x_1 = x_2$ and $y_1 = y_2$ **then**
- 16: Let $g_{est} \leftarrow g_1 + (g_4 - g_1)x - x_1/x_2 - x_1$
end
- 17: **if** $x_1 = x_2$ and $y_1 = y_2$ **then**
- 18: Let $g_{est} \leftarrow g_1 + (g_4 - g_1)y - y_1/y_2 - y_1$
end
- 19: **if** $x_1 = x_2$ and $y_1 = y_2$ **then**
- 20: Let $g_{est} \leftarrow g_1$
end
/
* The color difference between two regions is calculated using the
Gouraud distance. */
- 21: Let $Dc \leftarrow \|g - g_{est}\|_2$
/* Color differences are calculated using Euclidean distances. */

Output: Texture distance

3 Description of the NAMLab+ Algorithm

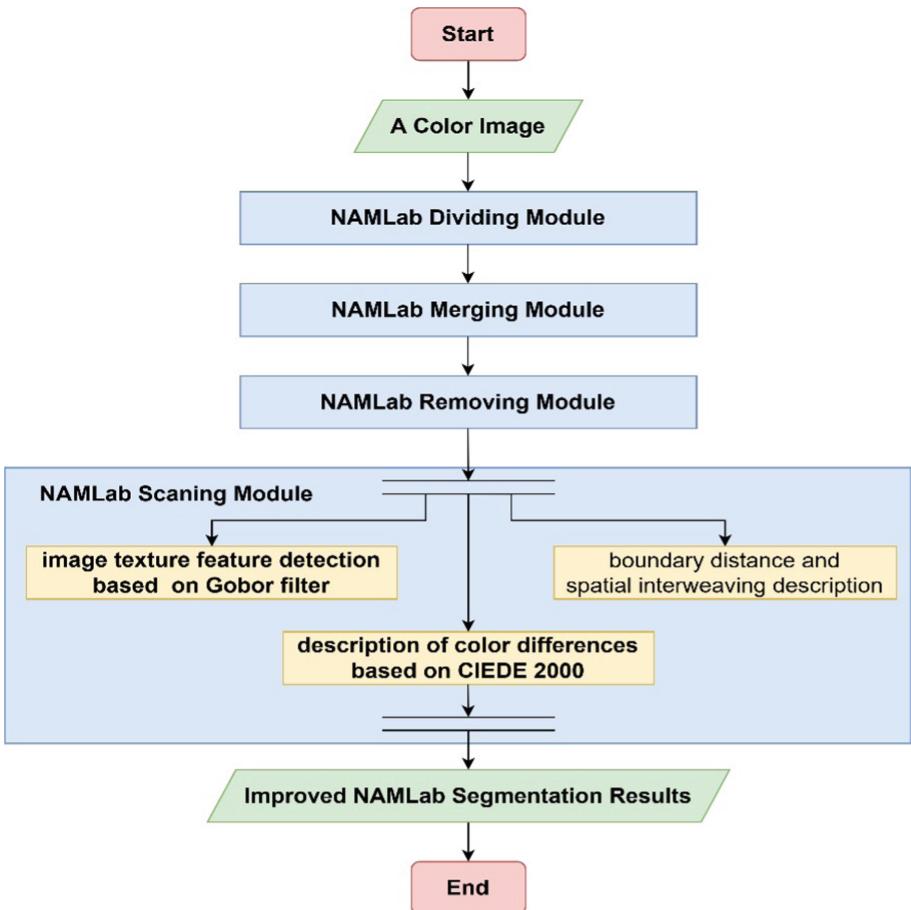


Fig. 1. The improved NAMLab algorithm flowchart

The NAMLab+ algorithm consists of four modules, the dividing module, the merging module, the removing module, and the scanning module. The dividing module is mainly based on the idea of the NAM. For an image, the Gouraud distance and the CIEDE2000 color difference formula are used to determine whether two pixels belong to the same NAMLab block, and the adjacent pixels continue to expand. Include the eligible pixels into the same block, otherwise create a new block, and finally divide the image into multiple NAMLab blocks of different sizes. A three-dimensional array is used to record the NAMLab block number of each pixel and the mean value of the Lab feature and variance. The time complexity of the NAMLab+ algorithm is the same as that of the NAMLab algorithm, and the time complexity of the two stages is better than that of the ICM algorithm.

The merging module needs to merge the divided NAMLab blocks correspondingly to prevent the phenomenon that the division granularity is too fine, thus affecting the final image representation effect. For two adjacent NAMLab blocks, the mean and variance of the Lab feature will be thresholded. If the difference between the mean and the variance of the two blocks is less than the threshold, the two NAMLab blocks can be merged. Starting from the left boundary, and we should scan and traverse the NAMLab block number to which each adjacent pixel belongs from bottom to top. If the adjacent pixel belongs to the different NAMLab block, the corresponding merging is performed according to the merging algorithm. Starting from the upper boundary, from left to right scan traverses the NAMLab block number to which each adjacent pixel belongs, and the rest of the processing methods are the same as above, until all adjacent pixels are processed.

The removing module is mainly in the process of NAMLab block merging. There will be a small amount of residual regions that cannot be merged. Therefore, in order to process the integrity of the image and the effect of image representation, the NAMLab+ algorithm calculates the residual region and the adjacent NAMLab region by calculating the residual region. The difference between the mean and the variance of the Lab eigenvalues, and the residual region is merged into the NAMLab region with the least difference.

The scanning module mainly performs the calculation of the relevant region indicators, including the calculation of the color difference between the regions, the calculation of the texture features between the regions, and the calculation of the boundary distance and spatial interweaving description.

The calculation of the color difference distance between the two regions is Eq. (7).

$$D_G(R_i, R_j) = \|G_{R_i} - G_{R_j}\|_2 \quad (7)$$

where G_{R_i} and G_{R_j} represent the texture feature values of the two regions R_i and R_j described by the Gabor filter, respectively.

The difference in color features is calculated by Eq. (8).

$$D_{AB}(R_i, R_j) = \|AB_{R_i} - AB_{R_j}\|_2 \quad (8)$$

where AB_{R_i} and AB_{R_j} represent the color feature value between the two regions.

The difference between the texture features of the two regions is calculated by Eq. (9).

$$D_T(R_i, R_j) = D_{AB}(R_i, R_j) * D_G(R_i, R_j) \quad (9)$$

When regions are merged, the boundary distance is calculated by Eq. (10).

$$D_B(R_i, R_j) = \frac{\sum_{p \in (R_i \cap B_{ij})} \sum_{q \in (R_j \cap B_{ij} \cap G_p)} \|C_p - C_q\|_2}{N_{pq}(R_i, R_j)} \quad (10)$$

In Eq. (10) B_{ij} denotes the border area between R_i and R_j , p and q denote pixels on the two sides of the border, which G_p are the local windows around p . At the same time, the window size distance is also introduced, which is described by Eq. (11).

$$D_N(R_i, R_j) = \frac{\min(N_{R_i}, ThNc)^{\frac{1}{t}} * \min(N_{R_j}, ThNc)^{\frac{1}{t}}}{\min(N_{R_i}, ThNc)^{\frac{1}{t}} + \min(N_{R_j}, ThNc)^{\frac{1}{t}}} \quad (11)$$

where N_{R_i} is R_i the total number of pixels in the region, N_{R_j} is the total number of pixels in the R_j region, t is an adjustable parameter, and $ThNc$ is the threshold value judgment as the window size. In the representation of the image, the same object may exist and the intensity changes. After merging and segmenting the region pixels, the image representation of the same object may form spatial interweaving in the spatial domain, which is described by Eq. (12).

$$D_I(R_i, R_j) = \min \left(\sum_{p \in R_i} f(M I_p, j), \sum_{q \in R_j} f(M I_q, i), \right) \quad (12)$$

The total region distance is calculated by Eq. (13).

$$D(R_i, R_j) = D_N(R_i, R_j) \frac{(\alpha D_M(R_i, R_j) + \beta D_T(R_i, R_j) + \gamma D_B(R_i, R_j))}{\sqrt{\lambda + D_I(R_i, R_j)}} \quad (13)$$

where α, β, γ , and λ represent the adjustable weight parameters of the color, the texture, the boundary information, and the spatial interleaving, respectively.

4 Experimental Results

The experimental results of the ICM, the NAMLab, and the NAMLab+ algorithms in this paper are run on an Intel(R) Core (TM) i7-9750H CPU with 2.60 GHZ. The capacity of the memory is 16 GB. To compare the efficiency and the quality of image segmentation, we compare the proposed NAMLab+ algorithm with state-of-the-art image segmentation algorithms, including the ICM algorithm, the gPb-OWT-UCM algorithm, and the NAMLab algorithm. The public datasets are BSDS300 [11], BSDS500 [11], MSRCv2 [22], VOC2012 [23], SBD [24], and NYUv2 [25], and the parameter metrics produced by each algorithm are evaluated on Matlab R2019a.

The image segmentation quality is currently mainly measured by CR [26], PRI [27], VI [28]. By setting the experimental values $K_L = 1.240$, $K_C = 1.040$, $K_H = 0.670$, $\alpha = 1.40$, $\beta = 1.54$, $\gamma = 1.82$, we test the proposed NAMLab+ algorithm on 100 test images on BSDS300 and 200 test images on BSDS500, respectively. Table 1 shows the number and the size of images on six popular datasets. Table 2 presents the results on BSDS300 and BSDS500. Table 3 shows the test results on other public datasets. During the experiments, we selected the optimal dataset scale (ODS) and the optimal image scale (OIS) [29, 30]. The former mainly uses fixed parameters to select the best for all images, while the latter is mainly used in each image to choose the best parameters based on the image. According to the three indicators of CR, PRI and VI on BSDS300, BSDS500, MSRCv2, SBD, VOC2012, and NYUv2, it can be seen that our NAMLab+ algorithm is very competitive with the NAMLab and the ICM algorithm.

Table 1. Dataset information

DataSet	Number of test image size	Average image size
BSDS300	100	444 × 358
BSDS500	200	428 × 374
MSRC	591	302 × 229
SBD	715	314 × 241
VOC2012	1449	470 × 386
NYUv2	1449	640 × 480

Table 2. Regional benchmarks for BSDS300 and BSDS500 datasets

Algorithm	BSDS300						BSDS500					
	CR		PRI		VI		CR		PRI		VI	
	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS
gPb-OWT-UCM	0.588	0.646	0.808	0.852	1.653	1.466	0.588	0.647	0.827	0.856	1.690	1.475
ICM	0.563	0.650	0.789	0.854	1.779	1.455	0.571	0.648	0.814	0.857	1.762	1.472
NAMLab	0.565	0.641	0.789	0.849	1.796	1.472	0.571	0.642	0.825	0.854	1.817	1.491
NAMLab+	0.571	0.640	0.792	0.849	1.794	1.484	0.566	0.636	0.807	0.852	1.821	1.511

Table 4 compares the time and the number of blocks of the two stages of the ICM algorithm, the NAMLab algorithm and the NAMLab+ algorithm on different data sets, and verifies the time complexity of the previous analysis. The time unit is the second in the Table 4. It can be seen that the total time on the dataset MSRCv2 for the ICM, the NAMLab, the NAMLab+ is 1.92 s, 1.07 s, and 0.93 s, respectively. The total time on the dataset SBD for the ICM, the NAMLab, the NAMLab+ is 2.28 s, 0.94 s, and 0.77 s, respectively. The total time on the dataset BSDS500 for the ICM, the NAMLab, the NAMLab+ is 3.20 s, 1.53 s, and 1.36 s, respectively. The total time on the dataset VOC2012 for the ICM, the NAMLab, the NAMLab+ is 4.03 s, 1.61 s, and 1.57 s, respectively. Therefore, on different datasets, the block time of our proposed NAMLab+ algorithm is even better than that of the NAMLab algorithm, and both are much better than the ICM algorithm.

In addition to the indicators to measure the segmentation quality, as shown in Fig. 2 we also selected five ground truth results drawn by experts, the UCM map of the gPb-OWT-UCM algorithm, the ICM algorithm, the NAMLab algorithm, and the NAMLab+ algorithm. Figure 2 (a), (b), (c), (d), (e), i.e., the first, the second, the third, the fourth, the fifth column of Fig. 2, are the ground truth results drawn by experts.

Table 3. Regional benchmarks for other datasets

Algorithm	CR		PRI		VI		Dataset
	ODS	OIS	ODS	OIS	ODS	OIS	
gPb-OWT-UCM	0.650	0.742	0.779	0.845	1.273	0.981	MSRCv2
ICM	0.642	0.748	0.748	0.854	1.205	0.958	
NAMLab	0.650	0.749	0.752	0.852	1.198	0.955	
NAMLab+	0.649	0.741	0.749	0.846	1.221	0.970	
gPb-OWT-UCM	0.582	0.641	0.862	0.892	1.877	1.617	SBD
ICM	0.581	0660	0.855	0.900	1.859	1.561	
NAMLab	0.571	0.653	0.852	0.897	1.885	1.578	
NAMLab+	0.567	0.648	0.850	0.895	1.913	1.611	
gPb-OWT-UCM	0.650	0.713	0.654	0.752	0972	0.922	VOC2012
ICM	0.670	0.737	0.669	0.780	0.935	0.790	
NAMLab	0.670	0.740	0.669	0.784	0.835	0.786	
NAMLab+	0.670	0.739	0.670	0.782	0.835	0.788	
gPb-OWT-UCM	Out of memory						NYUv2
ICM	0.453	0.502	0.845	0.862	2.592	2.346	
NAMLab	0.447	0.495	0.842	0.858	2.573	2.325	
NAMLab+	0.442	0.490	0.840	0.856	2.560	2.340	

Table 4. Blocking time, merging time and number of related blocks for each algorithm

Datasets	ICM				NAMLab				NAMLab+			
	Phase1	Seg Num	Phase2	Total	Phase1	Seg Num	Phase2	Total	Phase1	Seg Num	Phase2	Total
MSRCv2	0.79	1094	1.13	1.92	0.30	1504	0.77	1.07	0.38	1488	0.55	0.93
SBD	0.93	1065	1.35	2.28	0.23	1366	0.71	0.94	0.27	1379	0.50	0.77
BSDS500	1.53	1299	1.67	3.20	0.51	1473	1.02	1.53	0.60	1474	0.76	1.36
VOC2012	2.07	1177	1.96	4.03	0.67	1375	1.01	1.61	0.84	1361	0.73	1.57

Figure 2 (f), i.e., the sixth column of Fig. 2, is the UCM map of the gPb-OWT-UCM algorithm. Figure 2 (g), i.e., the seventh column of Fig. 2, is the UCM map of the ICM algorithm. Figure 2 (h), i.e., the eighth column of Fig. 2, is the UCM map of the NAMLab algorithm. Figure 2 (i), i.e., the ninth column of Fig. 2, is the UCM map of the NAMLab+ algorithm.

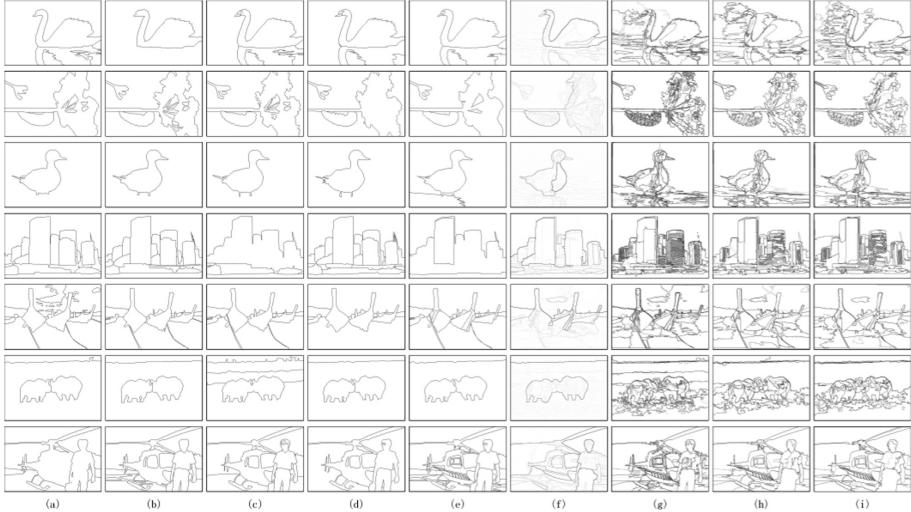


Fig. 2. Five ground-truth results with different resolutions on the BSDS500 dataset and three UCM maps.

When compared with the ICM algorithm and the NAMLab algorithm, it can be seen from the perspective of human vision that the NAMLab+ algorithm is more accurate for the outline description of the objects in the image. In addition, the amount of information of the object is also sufficient, which is more in line with human vision perception for our NAMLab+ algorithm.

Figure 3 presents the partial hierarchical image segmentation results on the BSDS500 dataset based on the ODS and the OIS over the CR. Figure 3 (a), i.e., the first column of Fig. 3, presents the original images. Figure 3 (b), (c), (d), i.e., the second, the third, the fourth column of Fig. 3, represent the segmentation results of the ICM algorithm, the NAMLab algorithm, and the NAMLab+ algorithm based on ODS over CR, respectively. Figure 3 (e), (f), (g), i.e., the fifth, the sixth, the seventh column of Fig. 3, represent the segmentation results of the ICM algorithm, the NAMLab algorithm, and the NAMLab+ algorithm based on OIS over CR, respectively.

It can be seen from the Fig. 3 that the ICM algorithm usually describes the region boundaries of objects in the image in excessive detail. From the perspective of human visual perception, the NAMLab+ algorithm can achieve better results than the NAMLab algorithm. The NAMLab algorithm fails to describe the original image well under ODS, while the NAMLab+ algorithm not only clearly describes the original image under the two indicators, but also accurately outlines the object boundary. It also has achieved good results in human visual perception.

As stated above, the experimental results in this section verify the theoretical our analyses.

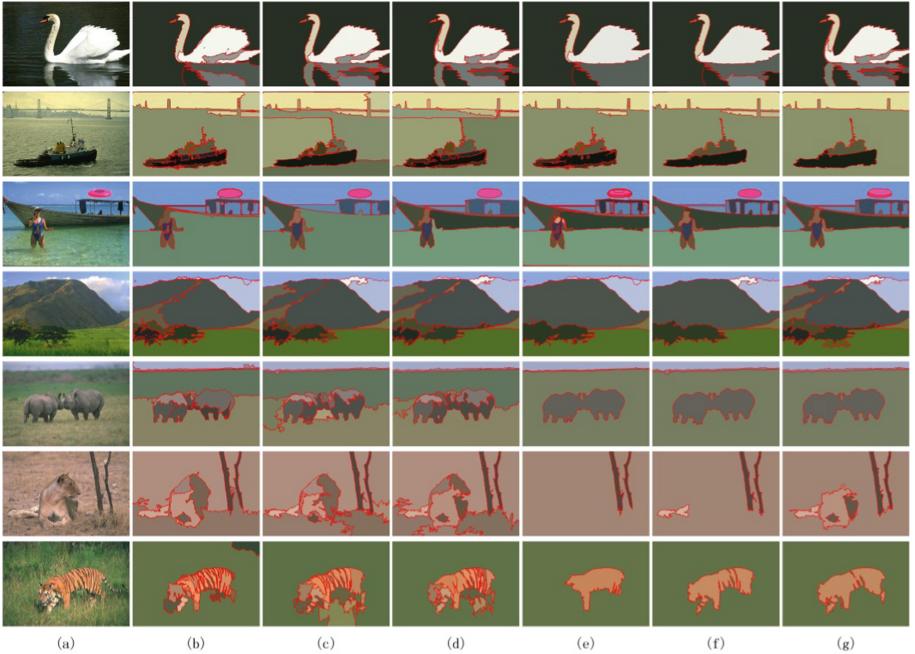


Fig. 3. Visual comparison of segmentation results produced by the ICM algorithm, the NAMLab algorithm and the NAMLab+ algorithm on the BSDS500 dataset.

5 Conclusion

In order to further improve the image segmentation quality and segmentation efficiency, in this paper, we improve the recently published NAMLab algorithm in two aspects. First, the CIEDE2000 color difference formula is used to replace the calculation formula of the color feature similarity in the original NAMLab algorithm and it is based on the human vision response to RGB and is used to accurately represent the reception of the color. Secondly, the calculation formula of texture features in the original NAMLab algorithm is modified. The improved algorithm is compared with the state-of-art algorithms in the field of image segmentation on six datasets, and it achieves better results in terms of visual presentation and the segmentation indicators.

Acknowledgment. This work is supported by the Natural Science Foundation of Guangdong Province of China under Grant No. 2017A030313349 and No. 2021A1515011517, and the National Natural Science Foundation of China under Grant No. 61300134, the National Undergraduate Innovative and Entrepreneurial Training Program under Grant No. 202110561070 and No.202110561066.

References

- Yan, Z., Zhang, J., Yang, Z., Tang, J.: Kapur's entropy for underwater multilevel thresholding image segmentation based on whale optimization algorithm. *IEEE Access* **9**, 41294–41319 (2021). <https://doi.org/10.1109/ACCESS.2020.3005452>
- Guo, R., Zhang, L., Yang, Z.: multiphase image segmentation model based on clustering algorithm. In: 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), pp. 1236–1239 (2021). <https://doi.org/10.1109/IPEC51340.2021.9421074>
- Bhandari, A., Singh, A., Kumar, I.V.: Spatial context energy curve-based multilevel 3-d Otsu algorithm for image segmentation. *IEEE Trans. Syst. Man Cybern. Syst.* **51**(5), 2760–2773 (2021). <https://doi.org/10.1109/TSMC.2019.2916876>
- Monemian, M., Rabbani, H.: Analysis of a novel segmentation algorithm for optical coherence tomography images based on pixels intensity correlations. *IEEE Trans. Instrum. Measur.* **70**, 1–12 (2021). <https://doi.org/10.1109/TIM.2020.3017037>
- Milano, F., Chevrier, A., De Crescenzo, G., Lavertu, M.: Robust segmentation-free algorithm for homogeneity quantification in images. *IEEE Trans. Image Process.* **30**, 5533–5544 (2021). <https://doi.org/10.1109/TIP.2021.3086053>
- Hussain, A., Khunteta, A.: Semantic segmentation of brain tumor from MRI images and SVM classification using GLCM features. In: Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 38–432020<https://doi.org/10.1109/ICI-RCA48905.2020.9183385>
- Özen, ŞK., Akşahin, M.F.: Automatic brain tissue segmentation on TOF MRA image. *Med. Technol. Congr. (TIPTEKNO)* **2020**, 1–4 (2020). <https://doi.org/10.1109/TIPTEKNO50054.2020.9299302>
- Khandelwal, M., Shirasagar, S., Rawat, P.: MRI image segmentation using thresholding with 3-class C-means clustering. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018, pp. 1369–1373 (2018). <https://doi.org/10.1109/ICSC.2018.8399032>
- Ilyasova, N., Shirokanov, A., Demin, N., Paringer, R.: Graph-based segmentation for diabetic macular edema selection in OCT images. In: 2019 5th International Conference on Frontiers of Signal Processing (ICFSP), pp. 77–81 (2019). <https://doi.org/10.1109/ICFSP48124.2019.8938047>
- Datta, A., Chakravorty, A.: Hyperspectral image segmentation using multi-dimensional histogram over principal component images. In: 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 857–862 (2018). <https://doi.org/10.1109/ICACCCN.2018.8748388>
- Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(5), 898–916 (2011)
- Syu, J.-H., Wang, S.-J., Wang, L.-C.: Hierarchical image segmentation based on iterative contraction and merging. *IEEE Trans. Image Process.* **26**(5), 2246–2260 (2017). <https://doi.org/10.1109/TIP.2017.2651395>
- Zheng, Y., Yang, B., Sarem, M.: Hierarchical image segmentation based on nonsymmetry and anti-packing pattern representation model. *IEEE Trans. Image Process.* **30**, 2408–2421 (2021)
- Luo, M.R., Cui, G., Rigg, B.: The development of the cie 2000 colour -difference formula: Ciede 2000. *Color Res. Appl.* **26**(5), 340–350 (2001)
- C. Gomez -Polo, MP Munoz, MCL Luengo, P. Vicente, P. Galindo, and AMM Casado, “Comparison of the cielab and ciede2000 color difference formulas,” *J. Prosthet. Dent.*, vol. 115, no. 1, p. 65 – 70, 2016

16. Zheng, Y., Yu, Z., You, J., Sarem, M.: A novel gray image representation using overlapping rectangular nam and extended shading approach. *J. Vis. Commun. Image Represent.* **23**(7), 972–983 (2012)
17. Liang, H., Zhao, S., Chen, C., Sarem, M.: The NAMlet transform: a novel image sparse representation method based on non-symmetry and anti-packing model. *Signal Process.* **137**, 251–263 (2017)
18. Zheng, Y., Sarem, M.: A fast region segmentation algorithm on compressed gray images using non-symmetry and anti-packing model and extended shading representation. *J. Vis. Commun. Image Represent.* **34**, 153–166 (2016)
19. Foley, J.D., Dam, A.V., Feiner, S.K., Hughes, J.F.: Computer Graphics, Principle, and Practice, 2nd edn. Addison Wesley, Reading (1990)
20. Wen, J., Zhisheng, Y., Hui L.: Segment the metallograph images using Gabor filter. In: Proceedings of ICSIPNN 1994. International Conference on Speech, Image Processing and Neural Networks, vol. 1, pp. 25–28 (1994). <https://doi.org/10.1109/SIPNN.1994.344974>
21. Dunn, D., Higgins, W.E.: Optimal Gabor filters for texture segmentation. *IEEE Trans. Image Process.* **4**(7), 947–964 (1995). <https://doi.org/10.1109/83.392336>
22. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: Proceedings of British Machine Vision Conference Coventry, UK, University of Warwick, September 2007, pp. 55.1–55.10 (2007). <https://doi.org/10.5244/C.21.55>
23. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)
24. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: Proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, Japan, September 2009, pp. 1–8 (2009)
25. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Proceedings of European Conference on Computer Vision, Firenze, Italy, October 2012, pp. 746–760 (2012)
26. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations, September 2007
27. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 929–944 (2007)
28. Meila, M.: Comparing clusterings by the variation of information. In: Schölkopf, B., Warmuth, M.K. (eds.) Learning Theory and Kernel Machines. LNCS (LNAI), vol. 2777, pp. 173–187. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45167-9_14
29. Syu, J.-H., S., Wang, S.-J., Wang, L.-C.: Hierarchical image segmentation based on iterative contraction and merging. *IEEE. Signal. Process. Soc.* **26**(5), 2246–2260 (2017)
30. Kim, T.H., Lee, K.M., Lee, S.U.: Learning full pairwise affinities for spectral segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(7), 1690–1703 (2013)



An Improved Block Truncation Coding Using Rectangular Non-symmetry and Anti-packing Model

Yunping Zheng^{1(✉)}, Yuan Xu¹, Jinjun Kuang¹, and Mudar Sarem^{2,3}

¹ School of Computer Science and Engineering, South China University of Technology,
Guangzhou, China

zhengyp@scut.edu.cn

² School of Software Engineering, Huazhong University of Science and Technology, Wuhan,
China

³ General Organization of Remote Sensing, Damascus, Syria

Abstract. Block truncated coding (BTC) is an image compression algorithm with a simple coding process and a high coding speed, which can be used in the field of military communications with high real-time requirements. As the price of pursuing simplicity and high speed, the compression ratio and the quality of the decoded image are sacrificed to some extent. Although some strategies have been proposed to improve the compression ratio and the quality of the decoded images, the effect is not obvious. Inspired by the quadtree-based block truncation coding (QEDBTC) and the non-symmetry and anti-packing model (NAM), in this paper, we propose a novel rectangular NAM-based block truncation algorithm (RNAMEDBTC), which uses rectangular NAM strategy to divide the initial blocks into rectangular homogeneous blocks. The spatial frequency measurement (SFM) is used as a measurement parameter to subdivide the initial blocks. For each homogeneous block, we replace the high and low quantization values and the binary bitmaps in the traditional block truncation coding with the average value of the pixels in the block, thereby a great improvement of the compression rate of the algorithm is achieved. In order to further improve the compression rate, we have increased the area of the smaller homogeneous blocks, and thus reducing the number of homogeneous blocks. These expanded blocks are called non-homogeneous blocks. For each non-homogeneous block, we need to do error diffusion block truncation coding (EDBTC) processing. The experimental results in this paper show that without degrading the quality of the decoded image, the proposed algorithm improves the compression rate significantly by 158.3% and 30.8% higher than the traditional BTC and QEDBTC algorithms, respectively.

Keywords: Block truncation coding · Non-symmetry and Anti-packing model · SFM

1 Introduction

Block truncation coding (BTC) [1] is a fast lossy image compression technique, first proposed by Delp and Mitchell in 1979. The biggest advantage of this algorithm was

that the encoding process was simple and had high-speed. This feature enables us to achieve high frame rate and high-resolution scenes of embedded monitoring systems under low-power processing constraints. It is often used in areas that require high real-time performance [2], and in the field of data hiding [3, 4]. At the same time, as the price of the block truncation coding for excessive pursuit of simple and high-speed algorithm, its compression rate and image quality have a lot of rooms for improvement.

In order to improve the compression ratio and the image quality, some strategies have been proposed in the recent years [5–9]. In the traditional BTC algorithm, all the pixel values in the block are represented by two high and low quantized values, which are calculated from the average and the standard deviation of the pixels in the block.

The absolute moment block truncation algorithm (AMBTC) proposed by Lema and Mitchell [10] improved the calculation method of high and low quantization values. It took the average value of some pixels whose pixel values were higher than the average value as the high quantization value, and took the average value of other pixels as a low quantization value. Error Diffused BTC (EDBTC) proposed by Guo [11] used error diffusion to diffuse the quantization errors to the adjacent pixels in order to maintain local average pixels. The resulting image showed good image quality and the block effect was reduced. But it required more processing time. Devi and Mathew proposed an image retrieval algorithm based on the EDBTC [12]. Later, Guo and Wu [13] proposed an ordered dither block truncation algorithm (ODBTC), which further increased the processing efficiency by using Look-Up-Table dither arrays. Guo and Liu [14] proposed a Dot-Diffused BTC (DDBTC) which maintained parallel processing capabilities while ensuring good image quality. Later Liu et al. [15] proposed an improved Near-Aperiodic DDBTC (NADDBTC), aimed to removing or reducing the degradation products of the periodic patterns and the impulsive noise. In addition to these halftone-based block truncation coding algorithms, Guo and Sankarasrinivasan proposed an improved multitone-based block truncation coding algorithm [16], and an image reconstruction algorithm based on multitone-block truncation coding [17]. In addition, Guo and Sankarasrinivasan have also established a block truncated coding image database based on halftone [18], which could be used in the field of deep learning. There were also some algorithms that focused on the block division method, such as the Quadtree-based EDBTC (QEDBTC) proposed by FJ Yang et al. [19], which used quadtree division method for dividing the initial blocks into sub-blocks of different sizes based on the global or the local SFM, and then these sub-blocks were encoded. With a certain increase in computational complexity, the image quality and the algorithm compression rate have been improved to a certain extent. However, due to its inherent overemphasis on symmetrical characteristics, the quaternary tree segmentation would cause some regions to be over-segmented and have too many blocks, which is not the most reasonable way to segment an image. The block truncation algorithm based on rectangular the non-symmetry and anti-packing model (NAM) which is proposed in this paper uses the rectangular NAM [20, 21] segmentation strategy for dividing the initial blocks, and uses the spatial frequency measurement (SFM) as the measurement parameter to subdivide the image into homogeneous blocks and non-homogeneous blocks. For the pixels in each homogeneous block, it is directly expressed by the average value of all pixels in the block; while for the non-homogeneous blocks, it needs to be processed by the EDBTC. The experimental results show that the

RNAMEDBTC proposed in this paper has a certain improvement in compression rate and the quality of the decoded image compared with the traditional BTC. Compared with the QEDBTC, the number of the divided blocks is less, and the compression speed and compression rate are higher, without reducing the image quality.

The rest of this paper is as follows. Section 2 briefly introduces several BTC algorithms. Section 3 details the RNAMEDBTC algorithm proposed in this paper. In Sect. 4, the experimental results are introduced and discussed. Finally, the conclusions are drawn in Sect. 5.

2 Block Truncation Coding (BTC) Algorithm

The main idea of the traditional BTC algorithm is to divide the input picture into non-overlapping blocks of a specified size, and then calculate the mean and the standard deviation of each block separately. Each block corresponds to a bitmap of the same size. For the pixel values greater than the mean, ‘1’ in the bitmap is written, otherwise ‘0’ in the bitmap is written, and then the mean, the standard deviation, and the bitmap are encoded. The decoding process is to calculate the high and the low quantization values based on the mean and the standard deviation, and replace ‘1’ and ‘0’ in the bitmap, respectively.

In the following sub-sections, we briefly introduce two BTC algorithms called Error Diffused Block Truncation Coding (EDBTC) and Quadtree-based Block Truncation Coding (QEDBTC) which are used for the comparison with our proposed RNAMEDBTC algorithm.

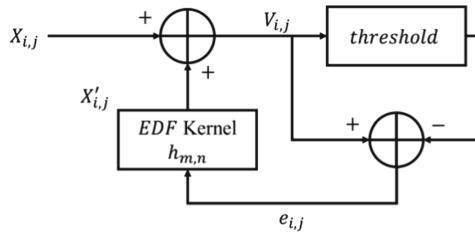


Fig. 1. Flow chart of error diffusion.

2.1 Error Diffused Block Truncation Coding (EDBTC)

The EDBTC proposed by Guo [11] was based on the traditional BTC algorithm, which added an error diffusion process to reduce the block effect of the BTC images without affecting the compression rate of the algorithm. The error diffusion process is shown in Fig. 1. Let's assume that the original image is first divided into multiple small blocks of size $n \times n$, and m is used to represent the total number of pixels in the small block, hence $m = n^2$. We calculate the average pixel value \bar{x} of each small block, the maximum pixel value x_{max} and the minimum pixel value x_{min} . Then, we do the error diffusion processing for each block, where $X_{i,j}$ represents the pixel value of the input coordinate (i, j) , $X'_{i,j}$

represents the error sum of the adjacent pixels of the coordinate (i, j) , $V_{i,j}$ represents the modified pixel value of the coordinate (i, j) , $O_{i,j}$ represents the binary output value at the coordinate (i, j) , $e_{i,j}$ represents the difference between $V_{i,j}$ and $O_{i,j}$, and $h_{m,n}$ represents the operator of the error diffusion.

The specific calculation formula is presented as follows:

$$V_{i,j} = +X'_{i,j}, \text{ where } X'_{i,j} = \sum \sum e_{i+m,j+n} \times h_{m,n} \quad (1)$$

$$e_{i,j} = V_{i,j} - O_{i,j}, \text{ where } O_{i,j} = \begin{cases} x_{\max} & \text{if } V_{i,j} \geq \bar{x} \\ x_{\min} & \text{if } V_{i,j} < \bar{x} \end{cases} \quad (2)$$

The commonly used error diffusion operator is Floyd operator which is given as follows:

$$\begin{bmatrix} - & * & 7 \\ 3 & 5 & 1 \end{bmatrix} \quad (3)$$

2.2 Quadtree-Based Block Truncation Coding (QEDBTC)

The QEDBTC algorithm was a new block truncation algorithm that Yang et al. [19] combined the quadtree segmentation with the EDBTC. The algorithm uses spatial frequency measurement (SFM) as a parameter to determine homogeneous blocks. And since this algorithm could calculate the SFM of the entire image and the local SFM of each small block, it was subdivided into two kinds: QEDBTC-GSFM and QEDBTC-LSFM. The approximate coding steps of this algorithm are as follows:

- (1) Divide the image I with size $2^N \times 2^N$ into multiple non-overlapping small blocks B of size $2^n \times 2^n$. The number of the small blocks is $m = 2^{(2N-2n)}$, where $B = \{B_0, B_1, \dots, B_i, \dots, B_{m-1}\}$ constructs 2^{2N-2n} queues $Q = \{Q_0, Q_1, \dots, Q_i, \dots, Q_{m-1}\}$, which used to store all the rectangular sub-patterns in each small block ($0 \leq i < m$).
- (2) Calculate the SFM. For QEDBTC-GSFM, the SFM of the entire image is calculated. As for QEDBTC-LSFM, the SFM of each small block is calculated.
- (3) Divide each B_i according to the quadtree segmentation method. The condition for stopping the quadtree segmentation is that the difference $diff$ between the maximum value V_{\max} and the minimum value V_{\min} of the pixels in the block is less than or equal to a threshold. The size of the threshold either depends on the SFM and the area of the block or reaches the minimum block of 4×4 .
- (4) If $diff$ is less than or equal to the threshold, the block is called a homogeneous block. When the EDBTC is processed for homogeneous blocks, only the *Indicator* representing the block size and the average pixel value in the block are stored in Q_i . If $diff$ is greater than the threshold, the block is called a complex block. Then, do EDBTC processing on the complex block to obtain a *bitplane*. When encoding, we need to store the *Indicator*, the V_{\max} and V_{\min} , and the *bitplane* in Q_i .
- (5) Finally, we repeat steps (3) and (4) to store the encoded results of all the small blocks in the queue Q .

Let M denote the number of rows of the block to be calculated, N denote the number of columns, R denote the row frequency, C denote the column frequency, and SFM denote the spatial frequency, then the formula for calculating the SFM is given as follows:

$$R = \sqrt{\frac{1}{MN} \sum_{m=1}^M \sum_{n=2}^N [x(m, n) - x(m, n-1)]^2} \quad (4)$$

$$C = \sqrt{\frac{1}{MN} \sum_{n=1}^N \sum_{m=2}^M [x(m, n) - x(m-1, n)]^2} \quad (5)$$

$$SFM = \sqrt{R^2 + C^2} \quad (6)$$

3 Methodology of Our Proposed RNAM-Based Block Truncation Method (RNAMEDBTC)

The RNAM-based block truncation algorithm (RNAMEDBTC) proposed in this paper is based on the QEDBTC, and it has achieved two improvements over the traditional QEDBTC:

- (1) It replaces the symmetric segmentation of the quadtree with the RNAM segmentation method which divides the small block into irregular rectangular blocks.
- (2) The corresponding relationship between the error parameter and the SFM has been changed. Also, the appropriate parameters are selected through the experiments.

3.1 Determination of Homogeneous Blocks and Selection of Parameters

The RNAMEDBTC algorithm needs to use the RNAM segmentation strategy in order to divide the initial block of the traditional BTC algorithm into two kinds of blocks, which are subdivided into homogeneous blocks and non-homogeneous blocks. The SFM is used as a measurement parameter for determining whether it is a homogeneous block or not.

V_{max} and V_{min} represent the maximum gray value and the minimum gray value respectively in the block to be determined. When the difference between the two values $diff = V_{max} - V_{min}$ is not greater than the threshold μ , it means that the block is a homogeneous block, otherwise, it is a non-homogeneous block. The parameter e_S used to expand the smaller rectangle is given the value 10. Similar to the QEDBTC, since the calculation range of the SFM is different, the algorithm can be subdivided into RNAMEDBTC-GSF and RNAMEDBTC-LSF. The correspondence between the threshold μ and the SFM in these two algorithms is also slightly different.

- (1) The formulas of the threshold μ and the SFM in RNAMEDBTC-GSF are given as follows:

$$\mu_{S \leq 16}^G = \begin{cases} 2.5 \times p, & SFM^G < 10 \\ 20 + 6 \times p, & SFM^G \geq 10 \end{cases} \quad (7)$$

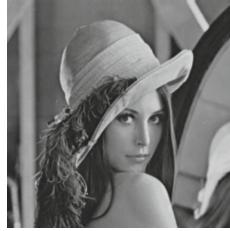


Fig. 2. Lena image of 512×512 .

$$\mu_{16 < S \leq 64}^G = \begin{cases} 2 \times p, & SFM^G < 10 \\ 20 + 4 \times p, & SFM^G \geq 10 \end{cases} \quad (8)$$

$$\mu_{64 < S \leq 196}^G = \begin{cases} 1.5 \times p, & SFM^G < 10 \\ 20 + 2 \times p, & SFM^G \geq 10 \end{cases} \quad (9)$$

Among them, the parameter S represents the area of the block, that is the total number of pixels. Regarding the choice of the parameter p , the method adopted in this article fixes a parameter p , and then changes the value of p to get a series of experimental data. The image used for testing is the Lena image of 512×512 , which is shown in Fig. 2. And the experimental results of using the QEDBTC-GSFM and the RNAMEDBTC-GSFM algorithms on Lena image under different p values are shown in Table 1.

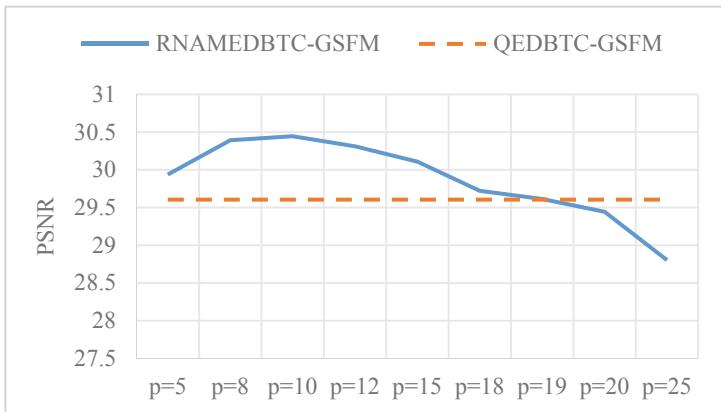


Fig. 3. $PSNR$ of QEDBTC-GSFM and RNAMEDBTC-GSFM algorithms under different p values

We have used the data in Table 1 to draw line charts for the $PSNR$ and the CR as shown in Fig. 3 and Fig. 4, respectively. For the RNAMEDBTC-GSFM algorithm, it can be analyzed from Fig. 3 that as the p value increases, the $PSNR$ first increases and then decreases, and the overall change is very small. Also for the RNAMEDBTC-GSFM algorithm, it can be seen from Fig. 4 that the compression

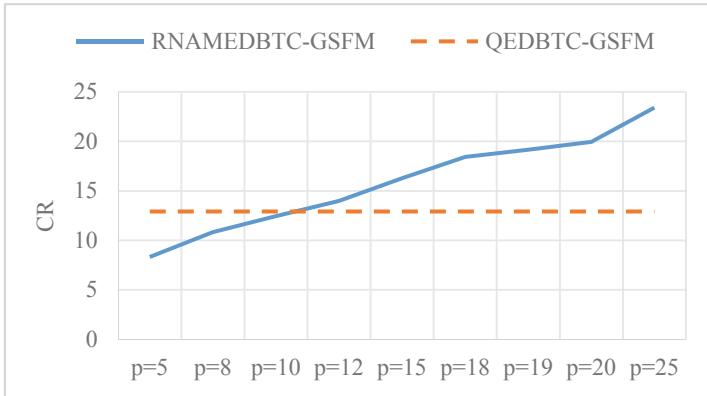


Fig. 4. CR of QEDBTC-GSFM and RNAMEDBTC-GSFM algorithms under different p values

ratio CR and compression speed are showing an increasing trend, and the number of blocks is showing a decreasing trend. When the p values are in the range [12, 19], the RNAMEDBTC-GSFM algorithm is better than the QEDBTC-GSFM algorithm either for the $PSNR$ or the CR . In this paper, we have chosen the middle value of the interval (i.e., $p = 15$) as an optimal appropriate parameter. At this value, compared with the QEDBTC-GSFM, the $PSNR$ increases by 0.5dB, the CR increases by 26%, the compression speed increases by 83%, and the Blocks decreases by 32%.

Table 1. Compression performance of QEDBTC-GSFM and RNAMEDBTC-GSFM algorithms on Lena under different p values.

		PSNR (dB)	CR	Time (ms)	Blocks
QEDBTC-GSFM		29.6047	12.91	3028	9319
RNAMEDBTC-GSFM	$p = 5$	29.9374	8.32	2510	12787
	$p = 8$	30.391	10.83	2135	10114
	$p = 10$	30.4447	12.44	1947	8955
	$p = 12$	30.3117	13.99	1803	8049
	$p = 15$	30.1082	16.27	1659	7059
	$p = 18$	29.7219	18.42	1567	6326
	$p = 19$	29.6141	19.15	1563	6119
	$p = 20$	29.444	19.95	1508	5898
	$p = 25$	28.8033	23.4	1398	5107

- (2) The formulas of the threshold and the SFM in the RNAMEDBTC-LSFM algorithm are given as follows:

$$\mu_{S \leq 16}^S = \begin{cases} 2.5 \times p_2, & SFM^S < 10 \\ p_1 + p_2 \times \left\lfloor \frac{SFM^S}{5} \right\rfloor, & 10 \leq SFM^S < 30 \\ 20 + 6 \times p_2, & SFM^S \geq 30 \end{cases} \quad (10)$$

$$\mu_{16 < S \leq 64}^S = \begin{cases} 2 \times p_2, & SFM^S < 10 \\ p_2 \times \left\lfloor \frac{SFM^S}{5} \right\rfloor, & 10 \leq SFM^S < 30 \\ 20 + 4 \times p_2, & SFM^S \geq 30 \end{cases} \quad (11)$$

$$\mu_{64 < S \leq 196}^S = \begin{cases} 1.5 \times p_2, & SFM^S < 10 \\ p_1 + p_1 \times \left\lfloor \frac{SFM^S}{5} \right\rfloor, & 10 \leq SFM^S < 30 \\ 20 + 2 \times p_2, & SFM^S \geq 30 \end{cases} \quad (12)$$

In these formulas, the selection of p_1 and p_2 is also done by using Lena image (Fig. 2) as a test image. The experimental results of multiple sets on different values of p_1 and p_2 show that when p_1 and p_2 are 20 and 16, respectively, the image quality is almost unchanged. However, the compression ratio and compression speed have improved significantly. The experimental results of using the QEDBTC-LSFM and the RNAMEDBTC-LSFM algorithms on Lena image when p_1 and p_2 are 20 and 16, respectively are shown in Table 2 below:

Table 2. Compression performance of RNAMEDBTC-LSFM and QEDBTC-LSFM on Lena

Heading level	PSNR	CR	Time (ms)	Blocks
RNAMEDBTC-LSFM ($p_1 = 20, p_2 = 16$)	29.6299	19.98	2073	6267
Increase (%)	0	23.9	-32.2	-45.8

As it can be seen from Table 2, when $p_1 = 20$ and $p_2 = 16$, the compression performances of the RNAMEDBTC-LSF algorithm for Lena image compared to the QEDBTC-LSFM algorithm are as follows. The *PSNR* is almost unchanged. The *CR* is increased by 23.9%. The compression speed increase by 32.2%, while the number of blocks is reduced by 45.8%.

3.2 RNAM Segmentation Algorithm

A particularly critical step in the RNAMEDBTC algorithm is the RNAM segmentation. The scanning method is to diagonally scan either to the right or to the downward. The rectangle with the largest area in the two scanning ways is adopted. The specific algorithm is given as follows:

- (1) Construct a marking matrix R with the same size as the block matrix to be divided.
- (2) Do raster scan the marking matrix and find the unmarked starting point as the upper left corner of the rectangle.
- (3) Start at the starting point and scan diagonally downward and to the right. If all points in the square are unmarked and they meet the judgment conditions of the homogeneous block, continue the diagonal scanning until the boundary is scanned. If there are marked points in the square, or the square does not meet the judgment conditions of the homogeneous block, stop the diagonal scanning, return to the previous point that meets the conditions, and record it as a square point.
- (4) Start with the square point and scan to the right. If all points within the rectangle formed by the point and the upper left corner of the rectangle are unmarked and they meet the homogeneous block judgment conditions, continue to scan to the right until the boundary is scanned. If there are marked points in rectangle, or the rectangle does not meet the judgment conditions of the homogeneous block, stop scanning to the right, return to the previous point that meets the condition. Record it as the lower right corner of the right rectangle, and calculate the rectangular area S_1 at this time.
- (5) Start with the square point and scan downward. If all points within the rectangle formed by the point and the upper left corner of the rectangle are unmarked and they meet the homogeneous block judgment conditions, continue to scan downward until the boundary is scanned. If there are marked points in the rectangle, or the rectangle no longer meets the homogeneous block judgment conditions, stop the downward scanning, return to the previous point that meets the condition. Record it as the lower right corner of the downward rectangle, and calculate the rectangular area S_2 at this time.
- (6) Compare the sizes of S_1 and S_2 , and take the rectangle with the largest area as the result rectangle.
- (7) Determine whether the area of the result rectangle is smaller than the minimum area parameter e_S or not. If it is not smaller, the *Indicator* assigns a value of ‘0’, indicating that the rectangle is a homogeneous block. If it is smaller, judge the type of the rectangle firstly, expand the lower right corner of the vertical rectangle to the right or expand the lower right corner of the horizontal rectangle to the downward until the area is not less than e_S and scan to the boundary. At this time, the rectangle is no longer a homogeneous block, and the *Indicator* is assigned a value of ‘1’. Mark all points corresponding to the rectangle in the marking matrix.
- (8) Repeat steps (2) to (7) until all pixels in the marking matrix R are marked, indicating that the block to be divided has completed RNAM segmentation.

3.3 RNAMEDBTC Algorithm Encoding and Decoding Processes

In the following two sub-sections, we introduce the steps for the encoding and decoding processes of our RNAMEDBTC algorithm in details respectively.

The steps of the encoding of the RNAMEDBTC algorithm are as follows

- (1) Divide the image into non-overlapping small blocks, where the size of each small blocks is $2^n \times 2^n$.
- (2) Calculate the *SFM*. Since the RNAMEDBTC algorithm is sub-divided into RNAMEDBTC-GSFM and RNAMEDBTC-LSFM, so if it is RNAMEDBTC-GSFM, the SFM of the entire image is calculated. If it is RNAMEDBTC-LSFM, the SFM of each small block is calculated.
- (3) Perform RNAM segmentation on each small block to find the largest homogeneous block rectangle that makes the difference *diff* between the maximum value V_{max} and the minimum value V_{min} of the pixels in the rectangular block less than or equal to a threshold μ , where the size of μ depends on the SFM and the block area. If the found rectangular area is less than the minimum allowable area e_S , it is appropriately expanded. The *Indicator* of this rectangular block is assigned a value of '1' indicating that it is a non-homogeneous block. If the found rectangular area is not less than the minimum allowable area e_S , then the *Indicator* is assigned a value of '0' indicating that it is a homogeneous block.
- (4) For homogeneous blocks with *Indicator* = 0, calculate the average gray value \bar{V} of all pixels in the rectangle, and perform *K*-code transformation [22] on the coordinates of the upper left corner of the rectangle (x_1, y_1) and the coordinates of the lower right corner (x_2, y_2) as follows: $k_1 = K(x_1, y_1)$, $k_2 = K(x_2, y_2)$, then calculate $\Delta k = k_2 - k_1$, and store \bar{V} and Δk into the rectangular sub-pattern queue Q .
- (5) For the non-homogeneous blocks with *Indicator* = 1, do EDBTC transformation to get the *bitplane*, calculate the median value: $V_{median} = \frac{(V_{max}+V_{min})}{2}$, and the difference value: $dev = V_{max} - V_{min}$. Build a *bitplane* of the same size as the rectangle, where the pixels which are greater than the average gray values \bar{V} are marked as '1' in the *bitplane*, otherwise they are marked as '0'. And store V_{median} , dev , and *bitplane* in the rectangular sub-pattern queue Q .

The *K* code transformation used in step (4) is defined as follows:

Suppose an image of $2^n \times 2^n$ size is represented by $F = \{f(x, y)\}$, where (x, y) is the coordinate expressed in binary as follows: $x = (x_{n-1}x_{n-2} \dots x_1x_0)_2$, $y = (y_{n-1}y_{n-2} \dots y_1y_0)_2$. Construct the variable k with the binary bits of x and y , so that $k = (y_{n-1}x_{n-1}y_{n-2}x_{n-2} \dots y_1x_1y_0x_0)_2$. This construction process is the positive transformation of the *K* code, denoted as $k = K(x, y)$. The *K* code transformation is a kind of dimensionality-reduction transformation from two-dimensional to one-dimensional, and its inverse process from one-dimensional to two-dimensional ascending-dimensional transformation is the inverse transformation of the *K* code denoted as $(x, y) = K^{-1}(k)$.

In the case of not using the *K* code transformation, the length and the width of the rectangular sub-pattern need to be stored, and in this case $2n$ bits are required. After using the *K* code transformation, only the difference between two coordinate *K* codes,

that is, Δk , with only n bits are required. And this can effectively reduce the number of the storage digits and increase the compression rate.

The steps of the decoding of the RNAMEDBTC algorithm are as follows

- (1) Construct a target matrix D of the same size as the original image, and assign all values to ‘−1’.
- (2) If the rectangular sub-pattern queue Q is not empty, take a rectangular sub-pattern R from Q , and do raster scan the target matrix D to find the first point with a value of ‘−1’ as the upper left corner coordinates of the decoded rectangle. However, if Q is empty, it indicates that the decoding is complete.
- (3) According to the number of the parameters in R , determine whether it is a homogeneous block or not. The rectangular sub-pattern of the homogeneous block contains two parameters \bar{V} and Δk , and the rectangular sub-pattern of the non-homogeneous block contains three parameters V_{median} , dev , and $bitplane$.
- (4) If the block is a homogeneous block, the inverse K code transformation is performed on Δk , and the coordinates of the lower right corner of the decoded rectangle are calculated according to the coordinates of the upper left corner. Then use the average gray value \bar{V} to replace the values of all points in D that form a rectangle with these two coordinates.
- (5) If the block is a non-homogeneous block, the coordinates of the lower right corner of the decoded rectangle are calculated according to the size of the $bitplane$, and then $V_{max} = V_{median} + dev$ and $V_{min} = V_{median} - dev$ are calculated according to V_{median} and dev . Then, use V_{max} to replace the value of ‘1’ in the $bitplane$ corresponding to the position of D , and use V_{min} to replace the value of ‘0’ in the $bitplane$ corresponding to the position of D .
- (6) Repeat steps (2) to (5) until Q is empty in order to complete the decoding process. Finally, calculate the $PSNR$ and the compression ratio CR .

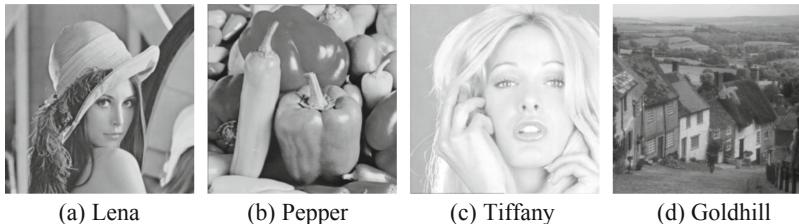


Fig. 5. Test images.

4 Experimental Results

In this section, we have used the $PSNR$, the CR , the compression time, and the *Blocks* to measure the performances of the compression algorithms. The compared algorithms are: the BTC, the AMBTC, the QEDBTC, and the RNAMEDBTC. The size of the initial division block of each of the four algorithms is 16×16 . The minimum area

parameter $e_S = 10$ in the RNAMEDBTC algorithm, in which the parameter $p = 15$ in the RNAMEDBTC-GSFM, and $p_1 = 20, p_2 = 16$ in the RNAMEDBTC-LSFM.

Assuming that the given image size is $2^N \times 2^N$, we have used $f(x, y)$ to represent the gray value at the coordinates (x, y) in the original image, and $g(x, y)$ to represent the gray value at the coordinates (x, y) in the decoded image, then the formula for calculating the *PSNR* is given as follows:

$$PSNR = 10\log_{10} \left\{ \frac{255^2 \times 2^N \times 2^N}{\sum_{x=0}^{2^N-1} \sum_{y=0}^{2^N-1} [f(x,y) - g(x,y)]^2} \right\} \quad (13)$$

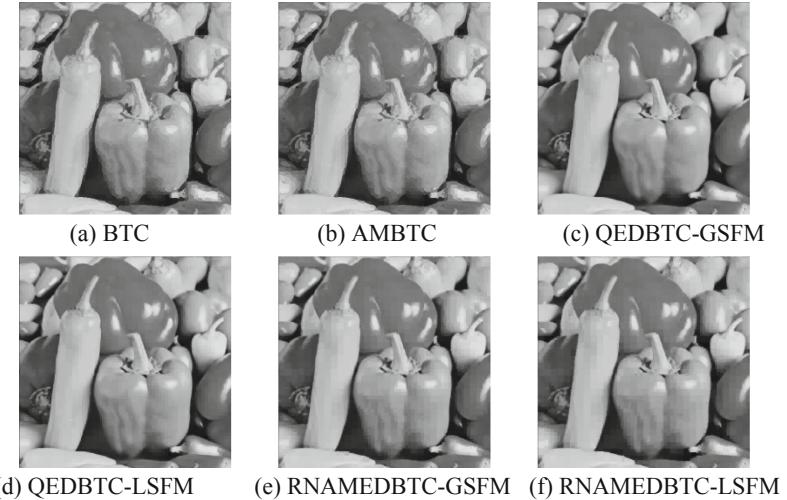


Fig. 6. Decoding images of pepper by using BTC, AMBTC, QEDBTC-GSFM, QEDBTC-LSFM, RNAMEDBTC-GSFM, and RNAMEDBTC-LSFM.

We have used B_S and B_C to represent the number of homogeneous blocks and the number of non-homogeneous blocks in the rectangular sub-pattern queue Q respectively. The statistical length of \bar{V} in each homogeneous block is 8 bits, and the statistical length of Δk is N bits. In each non-homogeneous block, the statistical lengths of V_{median} and dev are 8 bits and 4 bits respectively, and m and n represent the length and the width of the *bitplane*, respectively, and the formula for calculating the *CR* is given as follows:

$$CR = \frac{8 \times 2^N \times 2^N}{B_S \times (N+8) + B_C \times (m \times n + 8 + 4)} \quad (14)$$

The four grayscale images with a size of 512×512 used for testing are shown in Fig. 5. However, Fig. 6 shows the decoding images of the Pepper image resulting from the use of the BTC, the AMBTC, the QEDBTC-GSFM, the QEDBTC-LSFM, the RNAMEDBTC-GSFM, and the RNAMEDBTC-LSFM algorithms respectively. The compression performances of these algorithms are shown in Table 3.

It can be seen from Table 3 that the *PSNR* of the RNAMEDBTC algorithm proposed in this paper is larger than the *PSNR* of the BTC and AMBTC algorithms, and it is

Table 3. Compression performances of BTC, AMBTC, QEDBTC-GSFM, QEDBTC-LSFM, RNAMEDBTC-GSFM, and RNAMEDBTC-LSFM algorithms

	<i>PSNR (dB)</i>	<i>Time (ms)</i>	<i>Blocks</i>	<i>CR</i>
BTC	28.861	979	1024	7.53
AMBTC	29.219	912	1024	7.53
QEDBTC-GSFM	30.674	3659	9847	13.54
QEDBTC-LSFM	30.662	3270	9723	16.19
RNAMEDBTC-GSFM	30.78	2694	6716	17.66
RNAMEDBTC-LSFM	30.293	2243	6018	21.24

not much different from the QEDBTC algorithm, which is almost negligible. Since the principle of the algorithm proposed in this paper is to further divide the initial block of the traditional block truncation algorithm, so, both the compression time and the number of blocks will be larger than the traditional BTC and AMBTC algorithms. However, compared with the QEDBTC algorithm, the compression time of our RNAMEDBTC algorithm is reduced by 40.3%, and the number of blocks is also reduced by 53.7%. In terms of compression ratio, the compression ratio *CR* of the RNAMEDBTC algorithm has increased by an average of 158.3% compared to the BTC and the AMBTC algorithms, and by an average of 30.9% compared to the QEDBTC algorithm.

5 Conclusions

Inspired by the quadtree-based block truncation coding (QEDBTC) and the non-symmetry and anti-packing model (NAM), in this paper, we propose a novel rectangular NAM-based block truncation algorithm (RNAMEDBTC), which uses a rectangular NAM strategy to divide the initial blocks into rectangular homogeneous blocks. The experimental results in this paper show that without degrading the quality of the decoded image, the proposed algorithm improves the compression rate significantly by 158.3% and 30.8% higher than the traditional BTC and QEDBTC algorithms, respectively. These results verify the feasibility of the RNAMEDBTC algorithm, and also verify that the NAM segmentation is superior to the quadtree segmentation in block truncation algorithm.

Acknowledgement. This work is supported by the Natural Science Foundation of Guangdong Province of China under Grant No. 2017A030313349 and No. 2021A1515011517, the National Natural Science Foundation of China under Grant No. 61300134, and the National Undergraduate Innovative and Entrepreneurial Training Program under Grant No. 202110561070 and No.202110561066.

References

1. Delp, E., Mitchell, O.: Image compression using block truncation coding. *IEEE Trans. Commun.* **27**(9), 1335–1342 (1979)

2. Jiang, M., Yang, H.: Secure outsourcing algorithm of BTC feature extraction in cloud computing. *IEEE Access* **8**, 106958–106967 (2020)
3. Liu, X., Lin, C.C., Muhammad, K., et al.: Joint data hiding and compression scheme based on modified BTC and image inpainting. *IEEE Access* **7**, 116027–116037 (2019)
4. Shie, S., Jiang, J., Su, Y., Chang, W.: An improved steganographic scheme implemented on the compression domain of image using BTC and histogram modification. In: 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 640–644. Krakow (2018)
5. Cheng, H., Chen, C., Lee, L., Lin, T., Chiou, Y., Chen, S.: A low-complexity color image compression algorithm based on AMBTC. In: 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), pp. 1–2 (2019)
6. Guo, J., Sankarasrinivasan, S.: H-BTC database: a brief review on halftone based block truncation coding (H-BTC) images. In: 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 1–2 (2019)
7. Ke, S., Jhou, H., Chen, C., Lin, T., Abu, P., Chen, S.: A hardware-oriented image compression algorithm based on BTC and YEF color space. In: 2021 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–5 (2021)
8. Lamsrichan, P.: Straightforward Color image compression using true-mean multi-level block truncation coding. In: 2021 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–6 (2021)
9. Liao, J., Horng, J., Lee, C., Lu, H.: Hiding secret image in absolute moment block truncation code by using a block-selection scheme. In: 2019 8th International Conference on Innovation, Communication and Engineering (ICICE), pp. 39–42 (2019)
10. Lema, M., Mitchell, O.: Absolute moment block truncation coding and its application to color images. *IEEE Trans. Commun.* **32**(10), 0–1157 (1984)
11. Guo, J.: Improved block truncation coding using modified error diffusion. *Electron. Lett.* **44**(7), 462–464 (2008)
12. Devi, S., Mathew, A.: Fast image retrieval using error diffusion block truncation coding and unsupervised clustering. In: 2016 International Conference on Emerging Technological Trends (ICETT), pp. 1–6 (2016)
13. Guo, J., Wu, M.: Improved block truncation coding based on the void-and-cluster dithering approach. *IEEE Trans. Image Process.* **18**(1), 211–213 (2008)
14. Guo, J., Liu, Y.: Improved block truncation coding using optimized dot diffusion. *IEEE Trans Image Process* **23**(3), 1269–1275 (2014)
15. Liu, Y., Guo, J., Wu, Z., et al.: Near-aperiodic dot-diffused block truncation coding. *Signal Process.* **120**, 373–384 (2015)
16. Guo, J., Sankarasrinivasan, S.: Enhanced block truncation coding image using digital multitone screen. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 672–676. Kuala Lumpur (2017)
17. Guo, J.M., Sankarasrinivasan, S.: Reconstruction of multitone BTC images using conditional generative adversarial nets. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 814–817. Lanzhou, China (2019)
18. Guo, J., Sankarasrinivasan, S.: H-BTC database: a brief review on halftone based block truncation coding (H-BTC) images. In: 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 1–2. Taipei, Taiwan (2019)
19. Yang, F., Lien, C., Chen, P., et al.: An efficient quadtree-based block truncation coding for digital image compression. In: Thirtieth International Conference on Advanced Information NETWORKING and Applications Workshops, pp. 939–942. IEEE (2016)
20. Zheng, Y., Chen, C.: Study on a new algorithm for gray image representation. *Chin. J. Comput.* **33**(12), 2397–2406 (2010)

21. Zheng, Y., Yang, B., Sarem, M.: Hierarchical image segmentation based on nonsymmetry and anti-packing pattern representation model. *IEEE Trans. Image Process.* **30**, 2408–2421 (2021)
22. Zheng, Y., Chen, C.: A color image representation method based on non-symmetry and anti-packing model. *J. Software* **18**(11), 2932–2941 (2007)



Image Super-Resolution Reconstruction Based on MCA and ICA Denoising

Weiguo Yang¹, Bin Yang^{1(✉)}, Jing Li^{1(✉)}, and Zhongyu Sun²

¹ College of Information Science and Technology, Zaozhuang University, Zaozhuang 277160, China

batsi@126.com, jingl16233@163.com

² Qin Wei Middle School, Zaozhuang 277160, China

Abstract. Image super-resolution reconstruction is a high-resolution image that is reconstructed from a low-resolution image. The learning-based algorithm is one of the more effective algorithms for image super-resolution reconstruction, and the core idea of the algorithm is to use the sample library to train the information of the image in order to increase the high-frequency information of the test image and achieve the purpose of image super-resolution reconstruction. In this paper, we propose a new image super-resolution algorithm based on morphological component analysis and dictionary learning. Firstly we make independent component analysis for image denoising processing by the K-SVD method. And then, MCA algorithm is utilized to efficiently decompose low-resolution images into texture part and structure part. And the K-SVD method is used to make dictionary training of low-resolution images. The method not only improves the robustness of the images, but also adopts different reconstruction algorithms for the different characteristics of the texture and structure parts, which better retains the details of the images and improves the quality of the reconstructed images.

Keywords: Super resolution · Sparse representation · Dictionary training · Morphological layer segmentation analysis · Independent component analysis

1 Introduction

With the improvement of living standards, the demand for high-resolution images is increasingly urgent. In real life, limited by imaging equipment (such as cameras, camcorder), only blurry images can be obtained with very low resolution. However, clear high-resolution images are widely used in computer vision, medical images, video surveillance, and satellite imaging [1].

Since Tsai and Huang [2] first raised the issue of super-resolution reconstruction [3] in 1984, many methods of super-resolution reconstruction have emerged. It can be divided into three main categories: interpolation-based methods, reconstruction-based methods, and learning-based methods [4].

Learning-based algorithms focus more on the understanding of image content and structure than interpolation-based and reconstruction-based algorithms, utilizing more

priori knowledge on images. It is through the learning of high- and low-resolution images, establishing the relationship between them, and using this relationship as a priori information to provide stronger constraints, so that better results are often obtained.

Yang et al. [5] proposed a learning algorithm based on compressed sensing [6] to obtain high and low resolution dictionary pairs D_h and D_l by directly learning image libraries; obtain the relationship between high and low resolution images by learning high and low resolution dictionary pairs. The image quality of this algorithm reconstruction is better, but the influence of the trained sample is relatively large, the training speed is slow. The effect of the reconstruction depends more on the selection of the training sample, and does not consider the characteristics of the input image itself.

Jing et al. [7] proposed a modified algorithm based on Yang et al. [5], which first decomposes the low-rate image into two parts: texture and image, using the low-resolution texture method to get the high-resolution structural picture, and the structure and texture parts are added to get the final reconstructed picture. On the basis of Jing, MCA [8] is used to decompose low-resolution pictures. MCA decomposes texture and structure more thoroughly and can obtain the image features; and bicubic interpolation as an interpolation scheme can better recover high-resolution edge information.

In this paper, we propose a new super-resolution reconstruction algorithm based on MCA and dictionary learning. We first use ICA [9–12] for image denoising processing. Then, decompose low-resolution images into low-resolution texture images and structure images by MCA method. Finally, train low-resolution texture images to form an over-complete dictionary. The texture image contains complex information, super resolution reconstruction method based on sparse representation.

In the feature extraction process of the dictionary training stage, the second derivative is combined with the gradient direction, and in the process of dimensionality reduction, using the 2-Dimensional principal component analysis to reduce the dimensionality, and the dictionary trained by the K-SVD algorithm is used to reconstruct the texture image [13–15]. The structure image is relatively flat and can be obtained using the bicubic interpolation algorithm.

Finally, overlay the reconstructed texture image and the structural image to get the final high-resolution image. Experimental results show that compared with the traditional method and Jing's method, the proposed algorithm not only improves the convergence speed of the algorithm and the robustness of the image, but also improves the quality of the reconstructed image.

2 Methods

2.1 The ICA Basic Model

The problem of ICA $s(t) = [s_1(t), s_2(t), \dots, s_n(t)]^T$ can be described as follows. Suppose $x(t) = [x_1(t), x_2(t), \dots, x_m(t)]^T$ as the m -dimension observation signal vector, which is composed of a linear mixing of n unknown and independent source signals, where t is the discrete time and the value is as follows.

$$X(t) = AS(t) \quad (1)$$

where A is a $m \times n$ dimension matrix, called a hybrid matrix switcher.

The purpose of ICA is that in the case where the mixing matrix A and the source signal $s(t)$ are unknown, the separation matrix W is determined only according to the observation data vector $x(t)$, so that each output signal $y(t) = [y_1(t), y_2(t), \dots, y_n(t)]^T$ is defined as follows, which is an estimate of the source signal vector $s(t)$. And W is a $n \times m$ dimension matrix.

$$y(t) = Wx(t) = WAs(t) \quad (2)$$

2.2 Pre-processing of the Data

In general, the obtained data have correlations, so it is usually required to perform preliminary whitening or spherical processing of the data. Because the whitening process can remove the correlation between the observational signals, thereby the extraction process of subsequent independent components need to be simplified. In general, the algorithm with the data whitened converges better compared the one with the data whitening.

The random vector of a zero mean $z = (z_1, z_2, \dots, z_M)^T$ satisfies $E\{zz^T\} = I$, where I is the unit matrix. We call this vector the whitening vector. The essence of whitening is to remove correlations, which is the same as the goal of principal component analysis.

In the ICA, the independent source signals with zero mean $s(t) = [s_1(t), \dots, s_N(t)]^T$ have $E\{s_i s_j\} = E\{s_i\}E\{s_j\} = 0$, when $i \neq j$. And the covariance matrix is a unit matrix $\text{cov}(S) = I$ and the source signal $S(t)$ is white. For the observed signal, we should look for a linear transformation that is projected into a new subspace. For the observed signal $X(t)$, we should look for a linear transformation that makes $X(t)$ project into a new subspace and become a whitening vector, that is

$$Z(t) = W_0 X(t) \quad (3)$$

where, w_0 is the whitening matrix, and Z is the whitening vector.

Using the principal component analysis, we obtain a transformation by calculating the sample vector.

$$W_0 = \Lambda^{1/2} U^T \quad (4)$$

where U and Λ represent the eigenvector matrix and the eigenvalue matrix of the covariance matrix, respectively. It can be proved that the linear transformation W_0 meets the requirements of the whitening transformation.

This conventional method of whitening as a pretreatment of ICA can effectively reduce the complexity of the problem, and the algorithm is simple, which can be completed with traditional PCA. Pre-processing of the whitening of the observed signals with PCA enables the original solution mixing matrix is degenerated into an orthogonal matrix, which reduces the workload of ICA. In addition, PCA itself has a dimensionality reduction function, when the number of observed signals is greater than the number of source signals, the number of observed signals can be automatically reduced to the same as the number of source signal dimensions after whitening.

2.3 The ICA Denoising Process

The ICA can decompose the received mixed signals into independent components, and the separated components are the source signals. ICA has a good processing effect in denoising [16, 17], because it basically meets the premise requirements of ICA: noise and signal are independent of each other in time, and they synthesize to observe signals together. The steps for denoising with the ICA method are as follows:

- (1) Data acquisition;
- (2) ICA decomposition: using the FastICA [18–20] algorithm based on negentropy, x is the original signal acquired, W is the demixing matrix, separating the independent components y one by one;
- (3) Processing results: On the basis of (2), for the decomposed independent components, according to certain signal time domain and frequency domain and other priori knowledge, useful signals and noise signals can be identified, set the component of y that belongs to the noise zero, and then the x obtained $x = W^{-1}y$ is the original signal that removes the noise (Fig. 1).

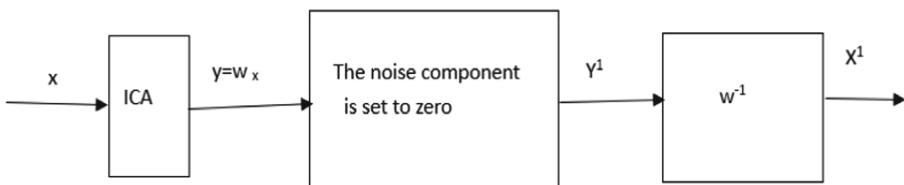


Fig. 1. Independent component denoising schematic 1.

The FastICA algorithm to estimate multiple components, we can calculate in the following steps:

1. Centralize the observed data X , so that its mean value is O ;
2. Whiten data $X \rightarrow Z$;
3. Select the number of components to be estimated, m , and set the number of iterations $p \leftarrow 1$;
4. Select an initial weight vector (random) W_p ;
5. Make $W_p = E\{Zg(W_p^T Z)\} - E\{g'(W_p^T Z)\}W$;
6. $W_p = W_p - \sum_{j=1}^{p-1} (W_p^T W_j) W_j$;
7. Make $W_p = W_p / \|W_p\|$;
8. If W_p does not converge, return to step 5;
9. Make $p = p + 1$, if $p \leq m$, return to step 4.

2.4 The MCA Image Decomposition Algorithm

The main idea of MCA is to use the morphological diversity of the different features contained in the image to give an optimal sparse representation of the image morphology.

MCA first extracts each morphological component of the signal separately according to the atoms in a given dictionary, and then looks for a solution to the signal decomposition inverse problem according to the sparsity constraint.

For a low-resolution image X with R pixels, MCA theory assumes that X is a linear combination of these two different parts: texture part X_t and structure part X_s ,

$$X = X_t + X_s \quad (5)$$

To separate low-resolution images X containing the texture part X_t and structure part X_s , MCA theory assumes that each part can be sparsely represented by a given dictionary, $D_t, D_s \in M^{R \times L}$ can be written as:

$$X_t = D_t \alpha_t, \quad (6)$$

$$X_s = D_s \alpha_s, \quad (7)$$

where α_t and α_s are the sparse representation coefficients of X_t and X_s in the corresponding dictionary D_t and D_s , for the low-resolution image X containing both the texture and structure parts, we need to find an optimum sparse representation through the dictionary D_t and D_s .

Optimum sparse representation of the low-resolution image X under the joint dictionary $\{D_t, D_s\}$:

$$\left\{ \alpha_t^{opt}, \alpha_s^{opt} \right\} = \arg \min_{\{\alpha_t, \alpha_s\}} \|\alpha_t\|_1 + \|\alpha_s\|_1 \text{ s.t. } X = D_t \alpha_t + D_s \alpha_s \quad (8)$$

3 Dictionary Training and Texture Image Reconstruction

Training dictionary is the most important step in image super-resolution reconstruction algorithms based on sparse representations. It will operate on the selected training library to train the dictionary corresponding to the high and low resolution. First, the second derivative is combined with the gradient direction in the feature extraction process to produce a new descent direction. An algorithm is designed with the new descending direction, which shows fast convergence speed and achieves better feature extraction results. Then dimensionality reduction in the dimension reduction process uses 2DPCA to eliminate the connection between rows and columns. Finally, complete the training with K-SVD.

3.1 2DPCA Reduces the Feature Dimension

The advantage of dimensionality reduction is energy saving in the subsequent computational training and super-resolution algorithm, before the dictionary learning reduce the dimensionality of the input low-resolution image block vectors, and the 2DPCA algorithm applied in these vectors, I expect to retain 99% of the average information on a subspace, while retaining 99% of the patches can be projected. The algorithm is as follows:

$m \times d$ Let the size of the image matrix A be $m \times n$, $X \in R^{n \times d}$ ($n \geq d$) as a matrix, its column vector is orthogonal to each other, after the linear transformation $Y = AX$, the image matrix A is projected to X, will produce the projection eigenvector Y. Optimum matrix X can be found by using the total measures of dispersion sample as a criterion function $J(X)$:

$$J(X) = \text{tr}(S_X) \quad (9)$$

where S_X is the covariance matrix of Y, $\text{tr}(S_X)$ for the trace of S_X .

$$\begin{aligned} J(X) &= \text{tr}\left\{E\left[(AX - E(AX))(AX - E(AX))^T\right]\right\} \\ &= \text{tr}\left\{X^T E\left[(A - EA)^T(A - EA)\right]X\right\} \end{aligned} \quad (10)$$

Then the image covariance matrix is defined as

$$G = E\left[(A - EA)^T(A - EA)\right] \quad (11)$$

Assuming that the number of training samples is M, matrix A_i ($i = 1, 2, \dots, M$), then the mean image is:

$$\bar{A} = \frac{1}{M} \sum_{i=1}^M A_i \quad (12)$$

Then the G is estimated as:

$$G = \frac{1}{M} \sum_{i=1}^M (A_i - \bar{A})^T (A_i - \bar{A}) \quad (13)$$

Make $X_{opt} = [X_1, X_2, \dots, X_d]$, X_{opt} is the optimum solution. After $X_{opt} = [X_1, X_2, \dots, X_d]$, feature extraction of the image, for the given A, $Y_m = AX_m$ ($m = 1, 2, \dots, d$).

This yields a set of later projected feature vectors, called the principal component vector of the image A.

From this, a set of projected eigenvectors $M = [Y_1, Y_2, \dots, Y_d]$ be obtained, which is called the principle component vector of image A.

3.2 K-SVD Dictionary Training

K-SVD dictionary training steps:

1. The high-resolution image library is under-sampling to obtain the corresponding low-resolution image library.
2. Extract the low-resolution image features. Images in the low-resolution image set were divided into $N \times N$ sized image blocks and features were extracted. The specific method is to use four one-dimensional filters:

$$f_1 = [-1, 0, 1], \quad f_2 = f_1^T \quad (14)$$

$$f_3 = [-1, 0, -2, 0, 1], \quad f_4 = f_3^T \quad (15)$$

where T represents the transposition. These four 1D filters are applied to the low-resolution image, so that each image block yields 4 eigenvectors which will be concatenated as a feature representation of the image block. Through high-pass filtering preprocessing, the gradient algorithm in the optimization method was improved. When $\frac{\partial^2 f}{\partial^2 x^2} \neq 0$, the second derivative and the gradient direction were combined to produce a new downward direction $d = \left[1 + \frac{\delta}{\frac{\partial^2 f}{\partial^2 x^2}} \right] \left(\frac{\partial f}{\partial x} \right)$. This method has fast convergence speed and better feature extraction effect.

3. Reduce the dimensionality of the resolution image with 2DPCA to train the low-resolution dictionary. Use the K-SVD algorithm train low-resolution image features into low-resolution dictionaries D_l .
4. Take the interpolated image set structure part. Interpolate the low-resolution training image to the same size as the high-resolution training image and decomposed it with MCA to obtain the structural part of the interpolated image.
5. Extract high-resolution image features. The remaining part of the high-resolution training image minus the low-resolution interpolated image structure part is taken as the texture part of the high-resolution image, and the texture part is divided into $(RN) \times (RN)$ sized image blocks and connected into vectors as eigenvector of the high-resolution image blocks.
6. Calculate a high-resolution dictionary. Assuming that high and low resolution image blocks have the same sparse representation coefficient α under high and low resolution dictionary pairs, the high resolution dictionary can be calculated by minimizing the lower formula approximation error:

$$D_h = \arg \min \|X_h - D_h \alpha\|_F^2 \quad (16)$$

Using Pseudo-Inverse:

$$D_h = X_h \alpha^+ = X_h \alpha^T (\alpha \alpha^T)^{-1} \quad (17)$$

where $+$ indicates a Pseudo-Inverse.

3.3 Redeling of Texture Images

Using the obtained D_l and D_h , low-resolution texture images can be reconstructed with high-resolution texture images. The low-resolution image is segmented according to $n \times n$ sized, and the two adjacent blocks overlap one pixel to make the corresponding adjacent high-resolution image blocks splicing smoother. The optimum sparse representation α of each block, makes $D_h \alpha$ represent the high-resolution image blocks, and this sparse representation can be solved by:

$$\min_{\alpha} \|\tilde{D} \alpha - \tilde{y}\|_2^2 + \lambda \|\alpha\|_1 \quad (18)$$

where $\tilde{D} = \begin{bmatrix} D_l \\ PD_h \end{bmatrix}$, $\tilde{y} = \begin{bmatrix} y \\ w \end{bmatrix}$, λ is the regularization coefficient, P is used to extract the overlapping area between the currently estimated high-resolution image feature block and its adjacent estimated feature block, and w represents the estimated value of the estimated high-resolution image feature block in the overlapping area. After obtaining the sparse representation α_i of each block, $D_h\alpha_i$ is the corresponding high-resolution image block, and all the high-resolution image blocks are stitched together to get the final high-resolution texture image.

4 Experiment and Result Analysis

The experimental data is single, which can prove the excellent performance of the method from different aspects, and can be expressed in different forms. The effect of the denoising method on the results of the scheme should be analyzed.

In this paper, the picture Lena is utilized to compare the proposed method with the traditional linear interpolation method and Jing's algorithm respectively. In the experiment, the regularization coefficient λ was 0.15, the image block size was 5×5 , 20000 image blocks were randomly selected for dictionary training, and the dictionary size was selected 256.

In this paper, image evaluation methods such as Peak Signal of Noise Ratio value and Structural Similarity Index Measurement value are used to evaluate the advantages and disadvantages of the reconstruction results. The results are shown in Table 1 and Table 2.

Table 1. PSNR values of different algorithms.

Image	Linear	Bicubic	Jing algorithm	Our algorithm
Lena	30.9799	32.7947	32.934	36.017
Barbara	26.5821	26.6594	26.80232	29.8862
Baboon	24.2037	24.6606	24.8462	27.8514

Table 2. SSIM values of different algorithms.

Image	Linear	Bicubic	Jing algorithm	Our algorithm
Lena	0.8601	0.8872	0.889	0.9012
Barbara	0.7812	0.79	0.8017	0.8219
Baboon	0.6188	0.6212	0.6377	0.6723

As can be seen from the table that the super-resolution reconstruction algorithm using MCA decomposition has improved PSNR and SSIM values from the traditional linear interpolation method and Jing algorithm. From Fig. 2, we can also intuitively see that the algorithm proposed in the paper has better results in details.



Fig. 2. The performances of Bicubic interpolation (left), Algorithm of Jing (middle) and our algorithm (right).

5 Conclusions

Text first uses ICA to Image Denoising, and applies MCA decomposition method to image super-resolution reconstruction based on sparse representation, improves the feature extraction and dimensionality reduction process of dictionary training, improves the convergence rate of the algorithm.

For the texture part and the structure part, the super-resolution reconstruction based on the sparse representation learning method and the bicubic interpolation are used respectively, which not only improves the robustness of the image, but also better preserves the detail information of the image, improves the quality of the reconstructed image, and achieves a better reconstruction effect.

However, the complexity of the algorithm is higher and the speed of the puzzle is slower, which increases the time for dictionary training and image reconstruction.

In the seasonal research, we will strive to find the algorithms with low complexity but good decomposition, or improve the MCA algorithm, so that it can reduce its algorithm complexity while ensuring the decomposition.

Acknowledgements. This work was supported by the talent project of "Qingtian Scholar" of Zaozhuang University, Youth Innovation Team of Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (No. 2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the PhD research startup foundation of Zaozhuang University (No.2014BS13), and Zaozhuang University Foundation (No. 2015YY02).

References

1. Chen, X., Qi, C.: Nonlinear neighbor embedding for single image super-resolution via kernel mapping. *Signal Process* **94**, 6–12 (2014)
2. Tsai, R.Y., Huang, T.S.: Multiframe image restoration and registration. *Adv. Comput. Vis. Image Process.* 317–339 (1984)
3. Park, S.C., Park, M.K., Kang, M.G.: Superresolution image reconstruction: a technical overview. *IEEE Signal Process. Mag.* **20**, 21–36 (2003)
4. Huang, D., Huang, W., Gu, P., et al.: Image super-resolution reconstruction based on regularization technique and guided filter. *Infrared Phys. Technol.* **83**, 103–113 (2017)

5. Yang, J., Wright, J., Huang, T., et al.: Image super resolution as sparse representation of raw image patches. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE Computer Society, Anchorage, AK, USA (2008)
6. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **19**(11), 2861–2873 (2010)
7. Jing, G., Shi, Y., Bing, L.: Single-image super-resolution based on decomposition and sparse representation. In: 2010 International Conference on Multimedia Communications, pp. 127–130. IEEE Computer Society, Hong Kong, China (2010)
8. Michael, E.: simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmon. Anal.* **19**(3), 340–358 (2005)
9. Zhang, Q., Yin, H., Allinson, N.M.: A simplified ICA based denoising method. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000), vol. 5479. IEEE Computer Society (2000)
10. Hyvarinen, A.: Survey on independent component analysis. *Neural Comput.* **2**, 94–128 (1999)
11. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. *Neural Netw.* **13**(4–5), 411–430 (2000)
12. Himberg, J. Hyvarinen, A.: Independent component analysis for binary data: An experimental study. In: Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001), pp. 552–556. San Diego, California (2001)
13. Jian, Y., David, Z., Frangi, A.F., et al.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 131–137 (2004)
14. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process* **54**(11), 4311–4322 (2006)
15. Rubinstein, R., Zibulevsky, M., Elad, M.: Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. *CS Technion* **40**(8), 1–15 (2008)
16. Barros, A., Mansour, A., Ohnishi, N.: Removing artifacts from electrocardiographic signals using independent component analysis. *Neurocomputing* **22**, 173–186 (1998)
17. Wisbeck, J.O., Barros, A.K., Ojeda, R.G.: Application of ICA in the separation of breathing artifacts in ECG signals. In: Proceedings of the International Conference on Neural Information Processing (ICONIP 1998), pp. 211–214. IOA publisher, Japan (1998)
18. Varinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*. Wiley, Wiley-Interscience Publication (2001)
19. Hyvarinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9**(7), 1483–1492 (1997)
20. Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Network* **10**(3), 626–634 (1999)



Image Representation Based on Overlapping Rectangular NAM and Binary Bit-Plane Decomposition

Yunping Zheng^{1(✉)}, Yuan Xu¹, Jinjun Kuang¹, and Mudar Sarem^{2,3}

¹ School of Computer Science and Engineering, South China University of Technology,
Guangzhou, China

zhengyp@scut.edu.cn

² School of Software Engineering, Huazhong University of Science and Technology, Wuhan,
China

³ General Organization of Remote Sensing, Damascus, Syria

Abstract. Binary-bit plane decomposition (BPD) is an effective method to reduce image complexity. Because of the diversity of sub-patterns of the non-symmetry and anti-packing model (NAM), combining NAM and BPD has resulted in a variety of image representation algorithms, such as a gray image representation based on square NAM and BPD (SNAMBPD), a gray image representation based on triangular NAM and BPD (TNAMBPD), and a color image representation based on rectangular NAM and BPD (RNAMBPD). The study found that the overlapping rectangular NAM has more advantages in image representation than the square NAM, triangular NAM and non-overlapping rectangular NAM. It can represent larger homogeneous blocks, that is, the number of homogeneous blocks obtained by using the overlapping rectangular NAM is much less. Based on this, this paper proposes a gray image representation based on overlapping rectangular NAM and BPD (ORNAMBPD), and extends a color image representation based on overlapping rectangular NAM and BPD (ORNAMBPD). The experimental results show that compared with LQT, SNAMBPD and TNAMBPD, our proposed ORNAMBPD has the least number of homogeneous blocks and the highest compression rate. Moreover, ORNAMBPD also has fewer homogeneous blocks than LQT and RNAMBPD algorithms, and the compression rate is also much higher.

Keywords: Image representation · Overlapping rectangular NAM (ORNAM) · Bit plane decomposition (BPD) · LQT

1 Introduction

Image representation is one of the research hotspots in the fields of image processing and pattern analysis [1–3]. The quaternary tree representation is the earliest form of hierarchical representation. Klinger [4] first proposed a method of expressing binary images with a quaternary tree. Later, in order to further reduce the amount of data required for storage, Gargantini [5] abolished the pointer and proposed a linear quaternary tree

representation method (LQT). The LQT can save 66% of storage space, and under special circumstances, it can even reach more than 90%. However, the LQT' method of segmenting homogeneous blocks emphasizes symmetry too much. In contrast, the non-symmetry and anti-packing model (NAM) proposed by Chen et al. [6] is more reasonable.

Due to the diversity of the NAM sub-patterns, a series of NAM-based image representation algorithms are proposed according to different types and combinations of sub-patterns. Using rectangular sub-patterns, Zheng et al. proposed a gray image representation based on the rectangular NAM [7]. After considering the overlap of rectangular sub-patterns, they proposed a gray image representation based on the overlapping rectangular NAM [8]. The rectangular sub-pattern is suitable for images with strong blocky structure, while for non-blocky images, the triangle sub-pattern can be used. Based on this, Zheng et al. proposed a gray image representation algorithm based on the TNAM [9], which is more efficient to process non-blocky images. Later, Fang et al. put forward an improved color image representation method by using direct non-symmetry and anti-packing model with triangles and rectangles [10]. By using the square sub-pattern, He et al. [11] proposed a square NAM representation method for binary images. Yi et al. [12] proposed a novel gray image representation method by using direct non-symmetry and anti-packing model with K-lines.

The NAM can also be combined with other representation algorithms to obtain some new image representation algorithms, such as a gray image representation based on the rectangular NAM and the extended Gouraud shading [13], an image representation based on the bit-plane decomposition and the triangular NAM (TNAMBPD) [14], an image representation based on bit plane decomposition and the square NAM (SNAMBPD) [15] and a color image representation based on the rectangular NAM and the bit plane decomposition (RNAMBPD) [16]. Inspired by the idea of the NAM in the Lab color space (NAMLab) and the “global-first” invariant perceptual theory, we proposed a novel framework for hierarchical image segmentation by taking a square sub-pattern as an example [17].

The ORNAMBPD algorithm proposed in this paper is inspired by the TNAMBPD and the SNAMBPD algorithms. First, a gray image is decomposed into eight binary images by using the BPD method, and then each binary image is divided into the overlapping rectangle homogeneous blocks, and finally the homogeneous blocks are coded. The main contributions of this paper are as follows.

Firstly, we propose an image representation based on the overlapping rectangular NAM and the binary bit-plane decomposition.

Secondly, another improvement of this algorithm is to determine when to encode the pixels with ones or zeroes, instead of uniformly encoding these pixels with ones.

Finally, on the basis of our proposed ORNAMBPD, a color image representation based on the overlapping rectangular NAM and the bit-plane decomposition (ORNAMBPD) is put forward.

The rest of this paper is as follows. Section 2 briefly introduces the BPD method. Section 3 presents our proposed ORNAMBPD algorithm in this paper. Section 4 further proposes an ORNAMBPD algorithm. In Sect. 5, the experimental results are discussed. Finally, the conclusions are drawn in Sect. 6.

2 Binary-Bit Plane Decomposition (BPD)

The idea of BPD is to divide a gray image into multiple binary images, which can greatly reduce the complexity of the image. The gray value range of a gray image is from 0 to 255, which can be represented by 8 bits. All pixel values of a gray image are represented by 8-bit binary, and the i^{th} bit of all pixel values is taken to form the i^{th} bit plane. Its formula is as follows:

$$V_{(x,y)} = \sum_{i=0}^{m-1} ai_{(x,y)} \times 2^i \quad (1)$$

$$I = \sum_{i=0}^{m-1} BP_i \times 2^i \quad (2)$$

where I represents the original gray image. BP_i represents the i^{th} binary image. $V_{(x,y)}$ represents the gray value at the coordinates (x,y) in I . $ai_{(x,y)}$ represents the value of the i^{th} bit in the binary representation of $V_{(x,y)}$ which is also the value at coordinates (x,y) in BP_i where $ai_{(x,y)} \in \{0, 1\}$. Taking Fig. 1 as an example, its bit-plane decomposition images are shown in Fig. 2.



Fig. 1. Pepper image.

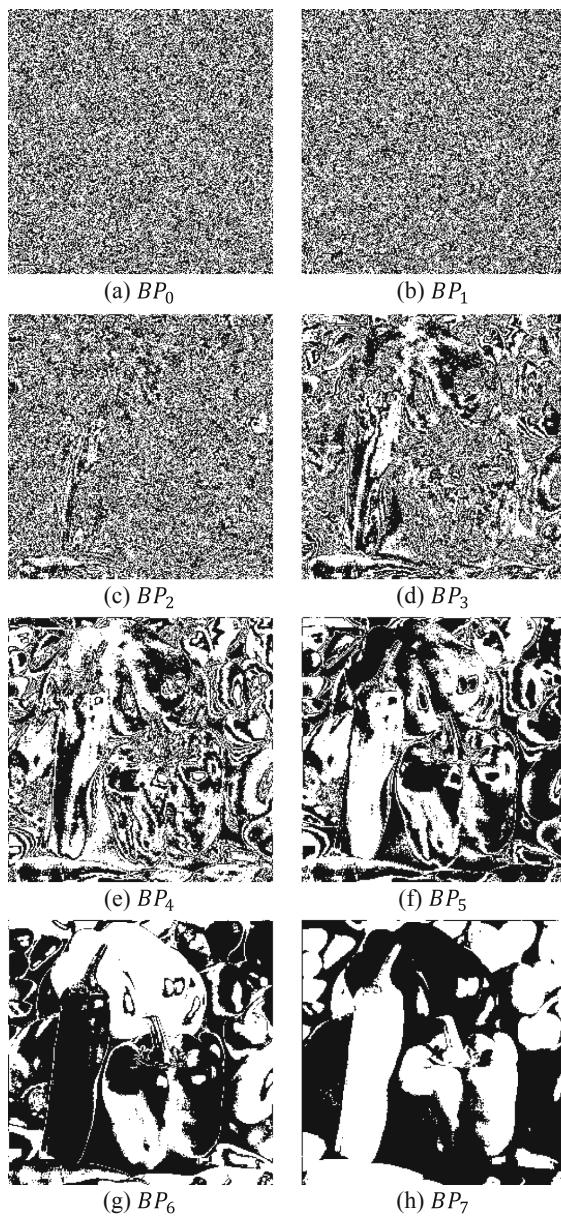


Fig. 2. Eight bit-plane decomposition images.

3 Proposed Gray Image Representation Based on Overlapping RNAM and Bit-Plane Decomposition (ORNAMBPD)

Compared with TNAMBPD and SNAMBPD, the ORNAMBPD algorithm proposed in this paper has two improvements:

(1) Replace the triangle sub-pattern in TNAMBPD and the square sub-pattern in SNAMBPD with overlapping rectangular sub-patterns. Line segments are regarded as a special case of rectangles. The comparison of these three sub-pattern layouts is shown in Fig. 3. It can be seen that the number of the overlapping rectangular sub-patterns (ORNAM) used for layout is less than those of the triangular sub-pattern (TRNAM) and the square sub-pattern (SNAM) layout.

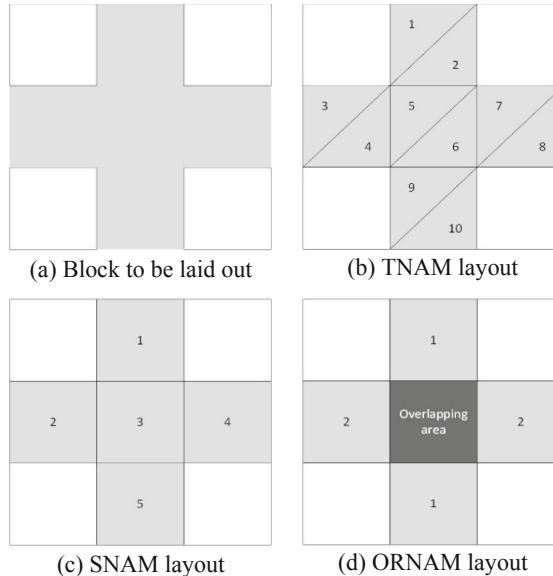


Fig. 3. Comparison of these three sub-pattern layouts.

(2) When encoding each binary image, both the TNAMBPD and SNAMBPD algorithms encode the point with a value of 1. The ORNAMBPD algorithm proposed in this paper first calculates the mean value V_{mean} of the binary image. If $V_{mean} > 0.5$, it means the number of points with a value of 1 in the binary image is large, choose to encode 0. If $V_{mean} \leq 0.5$, it means that the number of points with a value of 0 in the binary image is greater than or equal to the number of points with a value of 1, choose to encode 1.

The rest of this section will introduce the specific steps of the ORNAMBPD algorithm, the analysis of the ORNAMBPD segmentation algorithm and the analysis of storage structure of ORNAMBPD algorithm.

3.1 The Steps of the ORNAMBPD Algorithm

The coding steps of the ORNAMBPD algorithm are as follows.

Step 1. Suppose the size of the image I to be encoded is $2^n \times 2^n$, and the gray level is m . Construct a queue $CODES = \{BPCODE_0, BPCODE_1, \dots, BPCODE_m\}$ for storing the entire encoding result. $BPCODE_i = \{flag_i, Q_{ori}, Q_{pi}\}$, which is used to represent

the encoding result of the i^{th} binary image, where flag_i represents the pixel values of the i^{th} binary image. Q_{ori} and Q_{pi} are the queues for storing the overlapping rectangular sub-patterns and the isolated point sub-patterns in the i^{th} binary image, respectively. Decompose the image I into m binary images by using the BPD method to obtain $BP = \{BP_0, BP_1, BP_2, \dots, BP_m\}$.

Step 2. For each BP_i , first calculate the mean value V_{mean} . If $V_{\text{mean}} > 0.5$, then $\text{flag}_i = 0$, and code the point with the value 0. Otherwise, $\text{flag}_i = 1$, code the point with the value 1. The unmarked point in the BP_i with the pixel value of flag_i is regarded as the coordinates of the upper left corner of the overlapping rectangle which is denoted as (x_1, y_1) . Find the largest overlapping rectangle according to the ORNAM segmentation method which is put forward in the following Sect. 3.2, and perform the K -code transformation [8] on the upper left corner (x_1, y_1) and the lower right corner coordinates (x_2, y_2) of the rectangle. Work out $k_1 = K(x_1, y_1)$, $k_2 = K(x_2, y_2)$, store k_1 and k_2 into the queue Q_{ori} , and mark the points in the rectangle. If it is an isolated point, store k_1 of this point into the queue Q_{pi} and mark this point. Repeat the scanning process until all the points with the value of flag_i in BP_i are marked, and store flag_i , Q_{ori} and Q_{pi} into $BPCODE_i$.

Step 3. Go to Step 2 until the encoding of the entire gray image I is completed and obtain the queue $CODES$.

The decoding steps of the ORNAMBPD algorithm are as follows.

Step 1. Construct m binary images $BP = \{BP_0, BP_1, BP_2, \dots, BP_m\}$ with the size $2^n \times 2^n$, and a matrix MI with the size $2^n \times 2^n$ for storing decoded images;

Step 2. Extract $BPCODE_0, BPCODE_1, \dots, BPCODE_m$ from $CODES$ in turn. Extract $\text{flag}_i, Q_{\text{ori}}, Q_{\text{pi}}$ from $BPCODE_i$. If $\text{flag}_i = 0$, assign all BP_i to 1. If $\text{flag}_i = 1$, assign all BP_i to 0. Extract k_1 and k_2 from Q_{ori} , do K code inverse transformation, and obtain the coordinates of the upper left corner and lower right corner of the rectangle. Assign all the corresponding positions of the rectangle in the BP_i to flag_i . Extract k from Q_{pi} , do K code inverse transformation, and get the isolated Point coordinates. Assign flag_i to the corresponding position of the isolated point in BP_i . Repeat this process to assign values to all binary images.

Step 3. According to formula (2), use m binary images to synthesize the decoded image MI .

3.2 ORNAM Segmentation Algorithm

The steps of ORNAM segmentation algorithm are as follows.

Step 1. Judge the pixels to be coded according to the mean value V_{mean} of the binary image. If $V_{\text{mean}} > 0.5$ and $\text{flag} = 0$, encode pixels with value of 0. If $V_{\text{mean}} \leq 0.5$ and $\text{flag} = 1$, encode pixels with value of 1.

Step 2. Scan the block to be divided according to the horizontal priority and vertical priority strategy. Find an unmarked point (x_1, y_1) as the upper left corner of a rectangle and horizontally scan first. Then start to expand downward and stop when scanning to the boundary. Note that the coordinate of the point at this time are the coordinates (x_R, y_R) of the lower right corner of the horizontal priority strategy, and calculate the rectangular area S_R at this time.

Step 3. Start with the pixel in the upper left corner of a rectangle, and scan vertically first. Then start to extend to the right with this pixel until the boundary is scanned. At this time write down the coordinate of the pixel as the coordinate (x_C, y_C) of the lower right corner of the vertical priority strategy, and calculate the rectangular area S_C .

Step 4. Compare S_R and S_C , and take the rectangle with the largest area as the overlapping rectangle sub-pattern. If all the points in the overlapping rectangle sub-pattern are marked except for the upper left corner of the rectangle, then this point will be regarded as an isolated point. Just calculate $k_1 = K(x_1, y_1)$ and mark this isolated point. If there are other unmarked points in the rectangle in addition to the upper left corner of the rectangle, the coordinates of the upper left corner (x_1, y_1) and the lower right corner of the rectangle (x_2, y_2) are transformed by K code to obtain $k_1 = K(x_1, y_1), k_2 = K(x_2, y_2)$. Mark all points in the rectangle.

Step 5. Repeat Step 2 to Step 4 until all the points in the binary image are marked.

3.3 Analysis of Storage Structure of ORNAMBPD Algorithm

The ORNAMBPD algorithm first decomposes a gray image with a size of $2^n \times 2^n$ and a gray level of m into m binary images. The storage structure of a binary image encoding result $BPCODE_i$ is shown in Fig. 4. $flag_i$ is a flag value. The queue Q_{ori} that stores overlapping rectangles, and queue Q_{pi} that stores isolated points. Where $flag_i = 1$, it means that this binary image is encoding pixels whose value is 1; if $flag_i = 0$, it means that this binary image is encoding pixels whose value is 0.

$BOCODE_i :$	$flag_i$	Q_{ori}	Q_{pi}
--------------	----------	-----------	----------

Fig. 4. Storage structure of binary graph.

Q_{ori} stores the overlapping rectangles obtained by the ORNAM segmentation of the i -th binary image. An overlapping rectangle needs to store the K code of the upper left corner (x_1, y_1) and the lower right corner coordinates (x_2, y_2) of the rectangle: $k_1 = K(x_1, y_1), k_2 = K(x_2, y_2)$ (Figs. 5).



Fig. 5. Storage structure of an overlapping rectangle.

Q_{pi} stores the isolated points obtained by the ORNAM segmentation of the i -th binary image. An isolated point only needs to store the K code of the coordinate (x, y) : $k = K(x, y)$ (Figs. 6).



Fig. 6. Storage structure of an isolated point.

4 Proposed Color Image Representation Based on Overlapping RNAM and Bit-Plane Decomposition (ORNAMBPD)

The idea of the ORNAMBPD algorithm is to first divide the color image into three gray images, and then perform the ORNAMBPD algorithm for each gray image.

The coding steps of the ORNAMBPD algorithm are as follows:

- 1) Divide the color image MI with a size of $2^n \times 2^n \times 3$ and a gray level of m into three gray images MI_0, MI_1 , and MI_2 according to different color channels.
- 2) Decompose the gray image MI_i by BPD, $0 \leq i < 3$, and divide it into multiple binary images $BPI = \{BPI_0, BPI_1, BPI_2, \dots, BPI_{m-1}\}$, where $0 \leq j < m$.
- 3) Calculate the mean value V_{mean} of the binary image BPI_j in the gray image MI_i . If $V_{mean} > 0.5$, set $flag = 1$. Otherwise, $flag = 0$. Find the unmarked point (x_1, y_1) with the value of $flag$ as the starting point, which is also the upper left corner of the rectangle, and find the sub-pattern according to the ORNAM segmentation algorithm. If the sub-pattern is a rectangular sub-pattern, then the point (x_1, y_1) and the coordinates of the lower right corner of the rectangle (x_2, y_2) are transformed by K code: $k_1 = K(x_1, y_1)$, $k_2 = K(x_2, y_2)$. Store k_1 and k_2 into the queue $Qi_{or,j}$ which stores the rectangle sub-pattern parameters of the j^{th} binary image of the i -th gray image, and mark all the points in the rectangle. If the sub-pattern is the isolated point sub-pattern, perform K code transformation on the point (x_1, y_1) : $k_1 = K(x_1, y_1)$, store k_1 in the queue Qi_{pj} which is used to store the j^{th} binary image of the i^{th} gray image. These isolated points are marked.
- 4) If all the points with the value of $flag$ in the binary image BPI_j have been marked, store the $flag$, $Qi_{or,j}$ and Qi_{pj} into the queue Qi , which is used to store the encoding information of all the binary images obtained from the gray decomposition. Let $j = j + 1$. Go to step (3) until all the binary images of the gray MI_i have been processed. Store the queue Qi into the queue Q , which is used to store the Qi corresponding to all the gray images. Let $i = i + 1$. Go to step (3) until all gray images are processed.

- 5) Output queue Q .

The ORNAM segmentation algorithm mentioned in step (2) is the same as Sect. 3.2.

5 Experimental Results

5.1 Experimental Results of the ORNAMBPD Algorithm

Suppose a gray image with a size of $2^n \times 2^n$ and a gray level of m is processed by the ORNAMBPD algorithm. The number of overlapping rectangular sub-patterns generated by the i^{th} binary image is $N_{or}(i)$. The number of isolated point patterns generated is $N_p(i)$. The total amount of data after encoding the i^{th} binary image is $H_{BP}(i)$. The total amount of data H_{OR} that the gray image needs to store can be expressed as:

$$H_{OR} = \sum_{i=0}^{m-1} H_{BP}(i) = \sum_{i=0}^{m-1} [2nN_{or}(i) + nN_p(i) + 1] \quad (3)$$

If the LQT algorithm is used to process the gray image, each node needs $(3n - 1 + m)$ bits. N_{LQT} is used to represent the total number of nodes when represented by LQT, and $N_t(i)$ represents the number of triangle sub-patterns obtained by the TNAMBPD algorithm for the i^{th} binary image, $N_s(i)$ represents the number of square sub-patterns obtained by the SNAMBPD algorithm for the i^{th} binary image. ξ_{LQT_T} represents the ratio of the total amount of data that needs to be stored using the LQT algorithm to the total amount of data that needs to be stored using the TNAMBPD algorithm. ξ_{LQT_S} represents the total amount of data that needs to be stored using the LQT algorithm and the amount of data that needs to be stored using the SNAMBPD algorithm. $\xi_{LQT_{OR}}$ represents the ratio of the total data volume that needs to be stored using the LQT algorithm to the total data volume that needs to be stored using the ORNAMBPD algorithm. These formulas can be listed as follows:

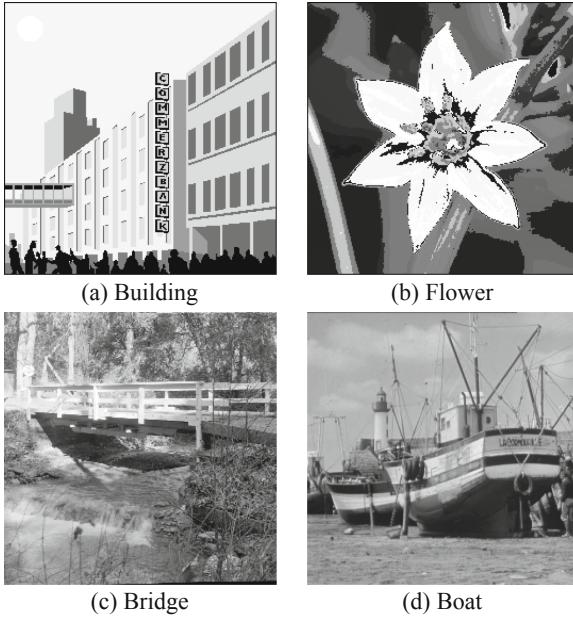
$$\xi_{LQT_T} = \frac{(3n-1+m)N_{LQT}}{\sum_{i=0}^{m-1} [(2n+1)N_t(i) + 2nN_l(i) + N_p(i)]} \quad (4)$$

$$\xi_{LQT_S} = \frac{(3n-1+m)N_{LQT}}{\sum_{i=0}^{m-1} [1.5nN_s(i) + 2nN_l(i) + N_p(i)]} \quad (5)$$

$$\xi_{LQT_{OR}} = \frac{(3n-1+m)N_{LQT}}{\sum_{i=0}^{m-1} [2nN_{or}(i) + nN_p(i) + 1]} \quad (6)$$

The four gray images with size 256×256 used to test the ORNAMBPD algorithm are shown in Fig. 7.

In Table 1, I represents the image used for testing. C represents complexity of the gray image. The calculation formula is $C = N_{LQT}/(2^n \times 2^n)$ [17], $(0 < C < = 1)$. $Block$ represents the number of the blocks divided by each algorithm.

**Fig. 7.** Test images for ORNAMBPD.**Table 1.** Comparison of the number of homogeneous blocks obtained by the four algorithms of LQT, TNAMBPD, SNAMBPD, and ORNAMBPD.

I	C	Block			
		LQT	TNAMBPD	SNAMBPD	ORNAMBPD
Building	0.2495	16351	11968	9577	3584
Flower	0.4589	30073	29193	26856	16096
Bridge	0.9910	64945	89625	86168	69433
Boat	0.9946	65182	90477	86833	71066
<i>Average value</i>		44138	55316	52359	40044

From Table 1, Table 2, and Table 3, the following conclusions can be drawn: The number of homogeneous blocks obtained by the ORNAMBPD algorithm is the least, especially when the complexity is low. Compared with the other three algorithms, the ORNAMBPD algorithm can even reduce by more than 62%. Compared with the LQT and TNAMBPD algorithms, the CR of the ORNAMBPD algorithm improves 1.65 times to 5.17 times, which is almost the same as the SNAMBPD algorithm. However, the total data amount required by the ORNAMBPD algorithm is the least when compared to the other three algorithms.

Table 2. Comparison of Compression ratio (CR) obtained by the four algorithms of LQT, TNAMBPD, SNAMBPD and ORNAMBPD.

I	C	CR			
		LQT	TNAMBPD	SNAMBPD	ORNAMBPD
Building	0.2495	1.0343	2.8219	10.1551	9.6164
Flower	0.4589	0.5624	1.2515	2.0578	2.3117
Bridge	0.991	0.2604	0.4298	0.4474	0.5734
Boat	0.9946	0.2595	0.4245	0.4326	0.5529
<i>Average value</i>		0.5292	1.2319	3.2732	3.2636

As stated above, it can be concluded that the ORNAMBPD algorithm proposed in this paper is superior to LQT, TNAMBPD and SNAMBPD. It also shows that the overlapping rectangular sub-pattern is more suitable for the gray image representation of the combination of BPD and NAM than the triangular sub-pattern and the square sub-pattern.

Table 3. The ratio of the total data volume of the LQT algorithm to the data volume of the three algorithms of TNAMBPD, SNAMBPD and ORNAMBPD.

I	C	ξ_{LQT_T}	ξ_{LQT_S}	$\xi_{LQT_{OR}}$
Building	0.2495	2.7282	9.8179	9.2972
Flower	0.4589	2.2253	3.659	4.1105
Bridge	0.991	1.6503	1.7182	2.2021
Boat	0.9946	1.6359	1.6674	2.131
<i>Average value</i>		2.0599	4.2156	4.4352

5.2 Experimental Results of the ORNAMBPD Algorithm

Assume that the size of a color image is $2^n \times 2^n \times 3$ and the gray level is m . Let the number of the overlapping rectangular sub-patterns generated by the j^{th} binary image of the i^{th} gray image is $N_{or}(i, j)$, and the number of the isolated point sub-patterns generated is $N_p(i, j)$, then the total data amount H_{ORC} that needs to be stored in the color image can be expressed as:

$$H_{ORC} = \sum_{i=0}^2 \sum_{j=0}^{m-1} [2nN_{or}(i, j) + nN_p(i, j) + 1] \quad (7)$$

The complexity of the color image is represented by C_c . Let N_{LQT} represent the number of nodes obtained when the LQT algorithm is used to represent the color image, then $C_C = N_{LQT}/(2^n \times 2^n \times 3)$.

The four color images with size 512×512 used to test the ORNAMBPD algorithm are shown in Fig. 8.



Fig. 8. Test images for ORNAMBPD

In Table 4, N_{LQT} , N_R , and N_{OR} represent the number of blocks obtained by LQT, RNAMBPD and ORNAMBPD algorithms, respectively. $D_{OR_{LQT}}$ represents the reduction rate of N_{OR} to N_{LQT} and is defined as $D_{OR_{LQT}} = (N_{LQT} - N_{OR})/N_{LQT}$. D_{OR_R} represents the reduction rate of N_{OR} to N_R and is defined $D_{OR_R} = (N_R - N_{OR})/N_{LQT}$.

Table 4. Comparison of the number of homogeneous blocks obtained by the three algorithms of LQT, RNAMBPD and ORNAMBPD.

I	Cc	N_{LQT}	N_R	N_{OR}	$D_{OR_{LQT}}(\%)$	$D_{OR_R}(\%)$
Flower	0.1953	153621	57078	52134	66.1	8.7
Gloriette	0.2216	174309	61140	57740	66.9	5.6
Flight	0.9119	717152	589685	523938	26.9	11.1
Lena	0.9862	775593	757957	729110	6	3.8
<i>Average value</i>		455169	366465	340731	25.1	7

In Table 5, C_{LQT} , C_R , and C_{OR} represent the compression ratios of LQT, RNAMBPD and ORNAMBPD algorithms, respectively. $R_{OR_{LQT}}$ represents the ratio

of the compression ratio of ORNAMBPD to LQT algorithm, which is defined as $R_{OR_{LQT}} = C_{OR}/C_{LQT}$. R_{OR_R} represents the ratio of compression ratio of ORNAMBPD to RNAMBPD algorithm, which is defined as $R_{OR_R} = C_{OR}/C_R$.

Table 5. Comparison of the number of CR obtained by the three algorithms of LQT, RNAMBPD and ORNAMBPD.

<i>I</i>	<i>Cc</i>	<i>C_{LQT}</i>	<i>C_R</i>	<i>C_{OR}</i>	<i>R_{OR_{LQT}}</i>	<i>R_{OR_R}</i>
Flower	0.1953	1.2	6.12	7.3	6.08	1.19
Gloriette	0.2216	1.06	5.72	7.02	6.62	1.23
Flight	0.9119	0.26	0.59	0.78	3	1.32
Lena	0.9862	0.24	0.46	0.57	2.38	1.24
<i>Average value</i>		0.69	3.22	3.92	5.68	1.22

From Table 4 and Table 5, it can be seen that the number of blocks obtained by the ORNAMBPD algorithm is reduced by an average of 25.1% than that of the LQT algorithm, and an average of 7% less than that of the RNAMBPD algorithm. In terms of compression rate, the algorithm proposed in this paper is also higher than LQT and RNAMBPD. The average compression rate of ORNAMBPD is 5.68 times that of LQT, which is 1.22 times that of RNAMBPD.

Therefore, our proposed ORNAMBPD algorithm for image representation is significantly superior to the LQT algorithm and the RNAMBPD algorithm for image representation.

6 Conclusions

Compared with the three gray image representation algorithms of LQT, TNAMBPD and SNAMBPD, the ORNAMBPD algorithm proposed by combining overlapping rectangular NAM and BPD in this paper produces fewer homogeneous blocks and higher compression rates. The color image representation algorithm ORNAMBPD, which extended the ORNAMBPD, is also significantly superior to the LQT algorithm and the RNAMBPD algorithm for color images. Therefore, image representation based on overlapping RNAM and bit-plane decomposition is a better image representation algorithm. Moreover, the overlapping rectangular NAM is also more suitable for the BPD-based image representation field than the triangle NAM, the square NAM and the non-overlapping rectangular NAM, and can be used as an effective supplement to the NAM-based image representation algorithms.

Acknowledgement. This work is supported by the Natural Science Foundation of Guangdong Province of China under Grant No. 2017A030313349 and No. 2021A1515011517, the National Natural Science Foundation of China under Grant No. 61300134, and the National Undergraduate Innovative and Entrepreneurial Training Program under Grant No. 202110561070 and No.202110561066.

References

1. Du, J., Li, W., Tan, H.: Three-layer image representation by an enhanced illumination-based image fusion method. *IEEE J. Biomed. Health Inform.* **24**(4), 1169–1179 (2020)
2. Wang, S., Ding, Z., Fu, Y.: Discerning feature supported encoder for image representation. *IEEE Trans. Image Process.* **28**(8), 3728–3738 (2019)
3. Zhang, S., Wang, J., Shi, W., Gong, Y., Xia, Y., Zhang, Y.: Normalized non-negative sparse encoder for fast image representation. *IEEE Trans. Circ. Syst. Video Technol.* **29**(7), 1962–1972 (2019)
4. Klinger, A.: Data structures and pattern recognition. In: Tou, J.T. (ed.) *Advances in Information Systems Science*. Springer, Boston (1978). https://doi.org/10.1007/978-1-4615-9056-9_5
5. Gargantini, I.: An effective way to represent quadtrees. *Comm. ACM* **25**(12), 905–910 (1982)
6. Chen, C., Zheng, Y., Sarem, M.: A novel non-symmetry and anti-packing model for image representation. *Chinese J. Electr.* **1**, 89–94 (2009)
7. Zheng, Y., Chen, C.: Study on a new algorithm for gray image representation. *Chinese J. Comput.* **33**(12), 2397–2406 (2011)
8. Zheng, Y., Chen, C., Li, Z.: Gray image representation algorithm based on overlapping RNAM. *J. Softw.* **23**(12), 3221–3232 (2012)
9. Zheng, Y., Chen, C., et al.: Study on an improved algorithm for TNAM of gray images. *J. Chinese Comput. Syst.* **30**(2), 322–326 (2009)
10. Fang, S., Chen, C., Zheng, Y.: An improved color image representation method by using direct non-symmetry and anti-packing model with triangles and rectangles. In: *Proceedings of JCAI'09*, pp. 440–443 (2009)
11. He, J., Zheng, Y., Guo, H.: A square NAM representation method for binary images. *Appl. Mech. Mater.* **143–144**, 755–759 (2012)
12. Yi, W., Xiao, R., Zheng, Y.: A novel gray image representation method by using direct non-symmetry and anti-packing model with K-lines. In: *Proceedings of JCAI'09*, pp. 476–479 (2009)
13. Zheng, Y., Sarem, M.: An improved gray image representation using overlapping rectangular non-symmetry and anti-packing and extended shading approach. In: *Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1863–1867 (2015)
14. Zheng, Y., Li, Z., Sarem, M., Wang, P., Hu, L.: An improved BPD-based triangle NAM for image representation. In: *Proceedings of ICNC 2011*, vol. 2, pp. 681–685 (2011)
15. Zheng, Y., He, J., Yang, Q., Xiong, Y.: A square non-symmetry and anti-packing model representation algorithm of gray images using binary bit-plane decomposition. In: *Proceedings of the 2015 11th International Conference on Natural Computation (ICNC 2015)*, pp. 934–938 (2015)
16. Zheng, Y., Chen, C.: A color image representation method based on non-symmetry and anti-packing model. *J. Softw.* **18**(11), 2932–2941 (2007)
17. Zheng, Y., Yang, B., Sarem, M.: Hierarchical image segmentation based on nonsymmetry and anti-packing pattern representation model. *IEEE Trans. Image Process.* **30**, 2408–2421 (2021)

Information Security



Research on the Rule of Law in Network Information Governance

Pei Zhaobin and Yu Yixiao^(✉)

College of Maritime Law and Humanities, Dalian Ocean University, Dalian 1160231, Liaoning,
China
13158970518@163.com

Abstract. Today, with the continuous development of globalization, Internet technology has achieved unprecedented leap-forward development. On the one hand, data in all fields of society is growing at an explosive rate, and the development of the Internet has brought about profound changes in the way of data transmission, storage and processing. The development of big data and related technologies enables the transmission and exchange of massive data and information to cross the limitations of time and space, and people's dependence on databases and information systems is also increasing. However, on the other hand, the development of the Internet is also a double-edged sword. For network users, once the data information on the network is leaked in large quantities, there will be unpredictable information security risks. In this context, the problem of network information security has become particularly serious, and the impact and harm to the entire society are very huge. It is necessary to strengthen the response, processing and management to fully guarantee the security of the Internet. In the existing legal governance system for network information security, our country's network security legislation has made rapid progress in recent years. In comparison, there is still a big gap. There are still many problems in China's network legislation, law enforcement entities, administrative supervision, governance strategies, and social law-abiding awareness, and it is impossible to effectively prevent various security problems that may occur in a comprehensive manner. This article will find out the problems we have in the legal governance of network information security from the actual situation in China. And by comparing and referring to the latest foreign research progress, it puts forward feasible suggestions and measures for the current China's network information security governance.

Keywords: Globalization · Internet development · Network information security · Rule of law governance

1 Status Quo of the Rule of Law in Network Information Security Governance

1.1 Status Quo of Domestic Governance

In the process of modernization of China's governance system and governance capacity under the background of globalization, network information security governance is an

important part of “China’s governance”. In recent years, the comprehensive information society has provided great convenience for people’s production and life, but its development also has some drawbacks, such as data theft, network attack, privacy leakage and other information security problems emerge in an endless stream. In the “overall national security concept” proposed by General Secretary Xi Jinping, the issue of network information security has been placed in a prominent position. This reflects our country’s emphasis on network information security, so it is very necessary to study network information security governance.

1.1.1 Chinese Laws and Regulations

Since China’s full-featured access to the Internet in 1994, our understanding of Internet governance has undergone a gradual process. Initially, people only regarded the Internet as a technical tool. “The Regulations on the Security Protection of Computer Information Systems” and “the Interim Provisions on the Management of the International Internet of Computer Information Networks” only treat the Internet as an emerging information technology. Since 2000, people have gradually realized the powerful media attributes, business opportunities and social value of the Internet. Regulations such as “the Telecommunications Regulations” and “the Measures for the Administration of Internet Information Services” have been promulgated, and various departments have begun to attach importance to participating in network information governance. At the same time, the “Electronic Signature Law” promotes the development of e-commerce, and the “Decision of the Standing Committee of the National People’s Congress on Maintaining Internet Security” emphasizes network security. With the in-depth popularization of Internet applications, the network has changed from a virtual space to an indispensable part of the real society, which has opened the stage of comprehensive socialization of China’s network. Some basic principles, principles and policies, laws and regulations and rules and regulations of network information security have been issued by China, which has gradually brought China’s information security on the right track. 2014 is the 20th anniversary of China’s full access to the international network. China is increasingly attaching importance to information security. The newly established Central Cyber Security and Information Committee of the Communist Party of China officially passed the “Internet Security Law of the People’s Republic of China” in 2016. China’s network information security. The rule of law system has been gradually improved and the level of security governance has been continuously improved, making China’s network information security governance a law-based system.

1.1.2 China’s Network Information Security Strategy

The promulgation of “Internet Security Law of the People’s Republic of China” and “the National Cyberspace Security Strategy” has pointed out the direction for China’s future cyberspace security governance. They both emphasize the defense of governance sovereignty. The former “focuses on the protection of domestic network security, and formally establishes the guiding position of the overall national security concept and the national security leadership system, which plays a crucial role in ensuring my country’s network information security [1]. “The latter analyzed the current network security

situation, identified “improving the network governance system” as one of the strategic tasks to ensure the smooth progress of network security work, and proposed that while maintaining national cyberspace security, it is also necessary to carry out international cooperation and strengthen international exchanges. Jointly build a harmonious network society, and comprehensively guarantee the peace and security of cyberspace [2]. Information security strategy is not only a long-term strategic deployment, but also a systematic project that adapts to the rapid development of contemporary information. Specifically, it refers to the whole process of strategic planning and implementation by a country in order to maintain its own overall security, solve the security threats of other countries in terms of information cyberspace, and make rational use of various national resources and technical means [3]. According to the definition of the national network information security strategy, the goal of my country’s network security strategy is to protect the security and interests of the country, maintain the stability of the cyberspace by enhancing the governance capability of the cyberspace to ensure the safety and feasibility of network activities, and thus promote national information and security. The healthy and long-term development of the information society [4].

1.1.3 The Characteristics of China’s Exploration of Network Information Security

Different from the US cyberspace strategy of “seeking security through strength”, China’s cyberspace strategy can be considered as “seeking security through governance”. Take comprehensive preventive measures to organize illegal intrusions, improve the level of national information network security governance, and advocate the implementation of the “active defense” governance model based on the “trinity”, that is, to comprehensively strengthen the information confrontation strength in cyberspace and protect user network information. Security, timely prevent various threats to information security, and safeguard national network information security and interests as a whole. At present, China’s network security governance has transformed from the traditional unified management and control mechanism to the modern network governance legal model, perfected the cyberspace legal governance structure, and basically formed a network information security governance legal system with Chinese characteristics.

1.2 Current Situation Abroad

In today’s world, North America represented by the United States is recognized as the region with the most developed information technology, while Europe is a region with a relatively open and complete information security system. In the Asian region where developing countries are concentrated, Japan, South Korea, Singapore and other developed countries are represented, and all countries regard the development of information technology as an important aspect of information security work. They attach importance to solving information security issues through legislation and other means. Taking a global view and improving information security management capabilities based on information opening has become the overall trend of the development of information security strategies in countries around the world. The research on the rule of law in foreign network information security governance started earlier and has reached the level

of national strategic management, which is ahead of China to a certain extent. Analyzing the development of foreign information security governance and discussing it for reference will help to improve the legal level of China's network information security governance.

1.2.1 European Union

ENISA, founded in 2004, formerly known as “Europe’s network and information security”, is a professional technology center, European support to member countries and organizations to develop a better ability of network security [5]. The telecommunications, judicial and intelligence departments of EU member states coordinate their work with each other, implement the EU and their own network security governance policies, and coordinate and cooperate with their own private sectors. At the same time, all countries set up a specialized agency of the network security, such as the national emergency response team, data and network security agencies responsible for monitoring dynamic network security in order to adjust the strategy. Since the Popularization of the Internet in Europe, the EU has formulated a series of laws and regulations to guide and regulate the development of the Internet both technically and administratively [6]. With the advancement of European integration, the EU has been abolishing old and outdated regulations and formulating new laws, gradually constructing and improving the legal framework of EU cyberspace, and guaranteeing the order and security of EU cyberspace. The EU has three major pieces of legislation on cyber security: Network and Information Security (NIS), General Data Protection Regulation (GDPR), and the EU Cyber security Act, which involves some significant issues such as how sharing authority between member states and the EU on national security issues can be divided [7].

1.2.2 The United States

In the United States, cyber security governance is carried out by the three parties, with internal government accountability, legislative and judicial bodies formulating relevant laws and regulations, and social accountability, forming a complete accountability system. The United States has a special government agency for Internet regulation, the Communications Commission. Its main responsibilities include: when there is a dispute over the division of Internet regulatory power, it has the right to formulate relevant laws to resolve the dispute. In addition, the implementation of the specific supervision of the Internet is divided by its authority. On the specialized management institutions have infrastructure protection committee, responsible for monitoring network infrastructure construction, to coordinate, network supervision and network information system updates, upgrades and technical protection. Federal statutes concerning network information security governance in the United States can be roughly divided into two categories: (1) Legal norms to prevent intrusions into computer systems, combat the manufacture and spread of computer viruses and malicious software, and protect information network infrastructure; (2) Legal norms to restrict and regulate the activities of network information release, dissemination and utilization [8]. Although the contents of relevant laws and regulations are overlapped and repeated, they completely cover

the regulation issues of network information and behaviors such as network infrastructure protection, network disclosure and data confidentiality, network terrorism, network pornography, network fraud, network intellectual property protection and so on.

1.2.3 Japanese

Japan's accountability system is similar to that of the United States. In addition to the government, legislative and judicial organs and the public, third-party accountability is added, which more obviously expresses the urgency of network security governance. Starting from its own economic and social development as well as the development of the information industry, Japan learns from the legislative characteristics of the Protection of network information security in Europe and the United States, and finally forms a legislative model combining unity and division. In terms of network information protection, Japan not only has a national information protection law, but also advocates that each industry in each region should formulate its own laws and regulations to protect network information security in accordance with its own reality [9]. In 2001, "the Basic Law on the Formation of a Highly Intelligent Communications Network society" was implemented. In 2014, "the Basic Law on Cyber Security" was implemented [10].

2 The Problems Existing in the Network Information Security of the Rule of Law in China

2.1 The Lack of Legislation

Since the 1990s, our country have issued a number of rules, regulations, methods of Internet information safety, such as "network security law", "computer information system safety protection regulations of the People's Republic of China", "computer virus prevention and control measures for the administration", "Internet network security measures for the implementation of information reporting," "Safety protection measures for the administration of international networking of computer information network", "the National People's Congress standing committee on maintaining Internet security", "Internet E-mail service management approach", "communication network security protection management approach", "telecommunications and the Internet users' personal information protection regulations", etc. These laws, regulations and measures provide corresponding provisions and norms for the activities of Internet actors from different sides. In today's society, Internet information security has aroused great attention from China's legislative circle. It can be said that China has established a basic framework in the legislation of Internet information security. But part of the laws and regulations, methods still remain in the level of words, specific practice has not been well implemented. For example, article 51 of the "Network Security Law" provides that the state establishes network security monitoring and early warning and information notification system. For safety detection and early warning and information notification, only the framework of early warning and notification system, as for the system of responsibility subject, monitoring mode, notification period are not clearly defined, lack of practical operability [11]. Although "network security Law" has become an important measure to protect network information security. In terms of liability punishment, although a

special chapter is set up to regulate the punishment of relevant illegal and criminal acts, from the content of the provisions, the punishment of most legal liability provisions is too light, mostly fines, too low illegal costs, which is not conducive to the realization of legislative purposes, so as to ensure the normal network order. However, because of the complexity of network security issues, the development of network technology is bound to bring new problems [12]. The introduction of “network security law” only represents the beginning of network security legislation, the future network security legislation can also make further improvement and efforts in the three directions of constructing an overall strategy to effectively protect network security in an open environment, actively participating in and promoting the international legislation process of network security, and establishing and improving the legal system of domestic network security.

2.2 The Unclear Rights and Responsibilities of Network Information Supervision Subjects

For a long time, China has a large number of departments with regulatory authority in the field of Internet supervision, which belongs to the joint supervision of multiple departments in the supervision mode, which is also the way of supervision in the traditional Internet period. However, in the face of the rapid development trend of mobile Internet, the supervision mode of the traditional Internet era still continues. Each department still acts independently and supervises according to the regulations of each department, and the relevant departments lack coordination ability and obvious lack of supervision basis. Therefore, the challenges facing China's network security legal system often come from the ambiguous rights and responsibilities of relevant departments and difficult coordination. On the one hand, China's vast legal system makes the distribution of rights and responsibilities among relevant departments unclear. Although the “Network Security Law” has clarified some responsibilities and rights of relevant departments in network security affairs, the responsibilities of network security protection and supervision and management of relevant departments are still not clear enough. On the other hand, China's cyber security legal system involves numerous legislative bodies and relevant departments, including the National People's Congress, The State Council, the Cyberspace Administration of China and local governments. So to a certain extent, led to the coordination between the law and operability has certain difficulty.

2.3 The Lack of Multi-dimensional Governance Strategy

At present, once there is a dispute case caused by violation of network security information, the public does not know any other way to protect their rights except to report to the public security department. However, the traditional public security department is not professional in dealing with the damage caused by network security exposure to the information subject in the mobile Internet environment, which may eventually lead to the failure of protecting their rights [13]. Although the government-led network security supervision mode can play a regulatory effect, it cannot fundamentally solve the network security problem. The lack of attention or excessive pursuit of economic interests within the industry is one of the main reasons leading to network information leakage. What's more, in today's fast changing network world, network technology development, rely

on the government the power to manage personal information leakage problem of the network world is not enough.

2.4 The Single Talent Training System

Although in the continuous development of globalization today, Internet technology has achieved unprecedented leap development. However, China's talent training model is still relatively simple, talent does not have comprehensive skills. Excellent network security personnel should not only have excellent professional skills but also have the corresponding legal knowledge. So cyber security talent remains a scarce resource across the country. Although the government has gradually solved this problem by setting up relevant majors in colleges and universities, the speed of network security personnel training in colleges and universities lags far behind the rapid development of the network security market, with a market gap of more than 90%. According to relevant analysis, China's demand for cyber security talents will reach 1.8 million by 2022, with only 20,000 cyber security graduates per year. Moreover, most training methods are more focused on technical training, and cannot be combined with the rule of law. This also leads to one of the reasons why government supervision is always difficult and ineffective when Internet information leakage occurs.

3 How to Improve the Rule of Law of Network Information Security in China

3.1 To Perfect the National Laws and Regulations

Developed countries such as Britain and the United States have introduced “computer abuse law”, “investigation power standard law” and other relatively sound network information security management legal system. Comparatively speaking, China still has bigger development space in information security governance legislation. We can gradually promote the improvement of relevant laws and regulations under the guidance of the overall national security concept, and eventually form laws and regulations that meet the needs of China's network information security governance. Therefore, China's network information security legislation should be based on the overall view of national security, in-depth exploration of information security problems and propagation path, the forefront of security technology research results as soon as possible with the rule of law integration, and under the framework of the rule of law integration and optimization of all aspects of social resources, so as to promote network information security governance technology, Institutions, laws and other aspects of an integrated, standardized pattern, to build a national information security legislation system with Chinese characteristics. The current legislative situation in China is that there are too few formal laws concerning the regulation of the mobile Internet industry, most of which are embodied in policies, notices and opinions. However, in the final analysis, these are not laws in the real sense and do not have the mandatory force of law. Therefore, it is urgent to vigorously accelerate the legislative process in the field of mobile Internet.

Cyber security legislation should be a long-term and systematic work. After the promulgation and implementation of the Basic Law on Cyber Security, relevant departments should, under the guidance of the basic Law's guiding ideology and legislative purposes, formulate legal norms at different levels in accordance with the actual needs of cyber security protection, and establish a targeted, systematic, operable and forward-looking system of cyber security laws and regulations [14]. The basic law determines the overall tone and core content of the entire network security legislation, reflecting the unity of principle and guidance; Other laws, administrative regulations and departmental rules mainly set specific norms for a certain aspect of network security, reflecting the applicability and effectiveness of laws. In supporting legislation, we should strengthen classified and itemized management, formulate special legislation in specific fields, such as "personal Network privacy Law", "network terrorism law" and other fields, formulate more detailed norms for such special fields, and realize the continuous development and improvement of network security legislation system.

3.2 Improve Network Law Enforcement Organizations

Referring to the specialized agencies established by the United States and the European Union, China can also learn from their practices and set up a specialized agency to coordinate and lead the government supervision of mobile Internet information security issues nationwide. Provinces and cities may set up their own network security supervision groups to undertake supervision responsibilities within their own regions. But it should also be subject to the supervision of national specialized regulatory agencies to form a unified and coordinated regulatory system. Change the chaotic situation of overlapping functions and powers of multiple departments.

Before the "September 11 Incident" in the United States, the government agencies in charge of network information supervision also had problems such as too many subjects, overlapping functions, and unclear powers and responsibilities to a certain extent. However, after the "9.11 Incident", especially after the establishment of the Department of Homeland Security, the United States has basically established a system coordinated by the President's Office of Critical Infrastructure Protection, with the Department of Homeland Security as the center, and the Department of Defense, Commerce, Administration and Management. The budget bureau and other agencies are supplemented by a relatively clear network management organizational system [15]. China can learn from the ideas of the United States in dealing with network information security. It is necessary to objectively evaluate the actual needs of network supervision and effectively integrate existing administrative regulatory resources, so as to reconstruct a set of administrative organization system with clearer levels and more unified powers and responsibilities, and comprehensively improve The overall strength of the Chinese government to prevent and respond to threats to network information security.

3.3 Establish Safety Accountability Mechanisms

Developed countries such as Russia, the United Kingdom, the United States and Japan have successively established cyber security accountability systems, which is a very necessary measure to mobilize the enthusiasm and responsibility of all sectors of society.

Let's take Japan as an example. The Japanese government's information security accountability system insists on the combination of promoting information disclosure and ensuring information security. At the same time, strengthen the improvement of information security technologies such as information screening and shielding, prohibit the transaction of information resources, and avoid information security crimes. The Japanese government has a relatively complete legal and regulatory system for information security accountability. Combining with the various information security regulations and strategic documents issued by the government under the concept of dynamic governance, and with the corresponding procedural laws, the scope and authority of information security policies have been greatly improved. The specific law enforcement work is actively carried out, so that legislation and law enforcement are closely linked. In addition, the legislative, judicial and administrative agencies have all created departments with information security accountability functions, and the public forces are flexibly involved in the whole process of accountability, which reflects the diversity and cooperation of the accountability bodies of the Japanese government, and makes the information security accountability effective. Furthermore, Japanese third-party organizations and enterprises have released information security status surveys, pointing out security risks and loopholes, and the public has boldly exposed organizations or individuals that may cause information security problems, so that Japan's information security is fully guaranteed. So China should carry out the overall construction from four aspects: government departments, legislative and judicial institutions, industry self-regulatory organizations and the public. We will develop a network information security accountability system and mechanism in line with China's actual situation.

3.4 Build a Comprehensive Talent Training System

In order to realize the transformation from science and technology to productivity, talent is an essential factor. The government must introduce supervision and technical talents with relevant legal knowledge matching the mobile Internet supervision environment, carry out the construction of professional talent team, and build a first-class team with both network technology and management level. From the perspective of professionals to ensure the improvement of professional technology, and finally use advanced technical means and efficient legal services to better protect the security of personal information. Talent is the carrier of technology, only with high-quality, highly skilled supervision team, to provide strong technical and regulatory support for the network according to the supervision and security protection, finally can escort personal information security.

3.5 Strengthen International Exchanges and Cooperation

As a big Internet country, China has been leading the world in the development of Internet information technology in recent years. However, the lack of legislation on network information security has seriously restricted the sustainable development of China's network security. In order to achieve the goal of cyber power, we should not only promote research on domestic cyber security legislation, but also actively promote international dialogue and research on cyber security issues, actively participate in the process of international cyber security legislation, master the right to speak in international cyber

legislation, and ensure the security of network information in an open environment. China should keep pace with the development of the Internet in the world with an active and open attitude, step up international cooperation in cyber security laws and regulations on the premise of guaranteeing cyber sovereignty, participate in the formulation of international Internet laws and regulations, and safeguard national sovereignty and interests in various forms [16].

Data flows were rare a dozen years ago. But now the data across borders surge is changing the dynamic development process of globalization [17]. The openness of the network determines that the legal regulation of network security cannot be completed by a single country. It is necessary for different countries and regions to strengthen communication and cooperation and develop unified standards in line with the development of network security. And the establishment of corresponding dialogue and coordination mechanism, as far as possible to ensure the orderly, coordinated and sustainable development of network security in the world [18]. As a major cyber country, While strengthening its own cyber security legislation, China should actively promote international dialogue and research on cyber security issues, actively participate in the process of international cyber security legislation, and contribute to the establishment of international dialogue mechanism on cyber security legislation. To show the image of a cyber power, enhance its voice in international Internet governance, and fully reflect its own interests.

4 Summary and Prospect

This paper summarizes the legal achievements and governance methods of network information security in China. On the basis of analyzing the legal system of network information security governance in China, this paper puts forward the problems existing in the legal system of network information security governance in China. For example, due to insufficient legislation, there are legal loopholes or even legal vacuum in the supervision process of relevant functional departments of the government, so that there is no supervision basis in the actual work process, resulting in difficult supervision work. Compared with the United States and other countries, there are some deficiencies in regulatory institutions and industry self-discipline mechanism. The supervision technology is backward and the supervision effect is poor. Personnel training mode is single, do not have comprehensive skills and other problems.

This paper draws lessons from foreign advanced experience and puts forward feasible suggestions for China's network information security governance. Although foreign mainstream information security accountability system, legal regulations, system framework and security standards have been basically formed, they all have their applicable conditions and scope, and even some defects need to be improved. In some aspects, it does not fully accord with the current situation and characteristics of China's network information security. Although cannot receive completely so, but can draw lessons from among them advanced method, standard and idea, form the network information security management system with Chinese characteristics.

In the face of the new situation of data security in the era of big data in the context of globalization, we must change our thinking, uphold the principle of equal emphasis on

development and security, and establish a multi-dimensional concept of risk prevention and control under the guidance of the overall national security concept [19]. With a profound vision to focus on the interests of citizens, societies and states behind technology. We will establish a legal mechanism for cyber security protection of critical infrastructure and improve the legal protection system for Internet crimes [20]. Strengthen the crackdown on network illegal and criminal acts, building a solid legal defense line for network information security.

References

1. Zhikai, Y., Yanfei, W.: New “National Security Law” Chinese information science in the background. *J. Inf.* **35**(7), 1–6 (2016)
2. Wei, H., Lin, Q., Liu Xiaoxin, X., Nuo, L.Y.: Research progress on network information security governance: Based on the current situation of the rule of law at home and abroad. *J. Intell.* **39**(4), 133–139 (2020)
3. Zhibin, H.: New security concept the theoretical construction of China’s network information security strategy. *J. Int. Observ.* **2**, 17–22 (2012)
4. Cuihong, C.: Comparison of Sino-US cyberspace strategy: goals, means and models. *J. Soc. Sci. Digest.* **3**, 8–10 (2019)
5. Regulation (EC): No 460/2004 of the European Parliament and of the Council of 10 March 2004 establishing the European Network and Information Security Agency [EB/OL]. Off. J. **L077**, 13/03/2004 P. 0001–0011, Recital 11
6. Alina Kaczorowska-Ireland. European Union Law. 4th edn. Routledge Cavendish, London, p. 176 (2016)
7. Mar Negreiro. ENISA and a new cybersecurity act, ERPS Briefing, February 26, 2019, [http://www.Europarl.europa.eu/RegData/etudes/BRIE/2017/614643/EPRS_BRI\(2017\)614643_EN.pdf](http://www.Europarl.europa.eu/RegData/etudes/BRIE/2017/614643/EPRS_BRI(2017)614643_EN.pdf)
8. Jianguo, Y.: The governance mechanism of network information security in the United States and its enlightenment to China. *J. Ref. Foreign Legal Syst.* **2**, 138–146 (2013)
9. Kuile, L.: Analysis on the characteristics of intelligence information sharing mechanism in the field of Japanese network security. *J. Intell. Explor.* **12**, 84–88 (2017)
10. Shuyi, W.: Japanese cyber security strategy: development characteristics and reference. *J. Chin. Adm.* **1**, 152–156 (2015)
11. Xie, Y.J., Jiang, S.L.: Analysis of the situation and problem on the legislation of cyberspace in China. *Chinese J. Netw. Inf. Secur.* **1**, 24–30 (2015)
12. Yuxiao, L., Wu, H., Yongjiang, X.: On the perfection of China’s network security legal system. *J. Eng. Sci.* **6**, 30 (2018)
13. Cuihong, C.: The cyberspace strategy is: target, method and model. *J. Soc. Sci. Abst.* **3**, 8–10 (2019)
14. Jon, R., Lindsay, T.M., Cheung, D.S.: Reveron:China and Cybersecurity: Espionage, Strategy, and Politics in the Digital Domain. Oxford University Press, Oxford (2015)
15. Hongwei, G., Jiahang, Y., Linjie, T.: American Federal Government information security accountability system and its reference. *J. Inf. Theory Pract.* **41**(8), 149–153 (2018)
16. Mergenthaler, S.: Managing Global Challenges: The European Union, China and EU Network Diplomacy, Springer Fachmedien Wiesbaden (2015)
17. McKinsey Global Institute. Digital Globalization: The New Era of Global Flow, 3 (2016)
18. Hanyang, C.: Research on network security analysis based on big data technology. *J. Netw. Secur. Technol. Appl.* **8**, 61–63 (2021)

19. Bhasin, M.: Challenge of guarding online privacy:role of privacy seals, government regulations and technological solutions. *J. Socio-Econ. Probl. State* **15**(2), 85–91 (2016)
20. Dunbar, D., Proeve, M., Roberts, R.: Problematic Internet usage self-control dilemmas: the opposite effects of commitment and progress framing cues on perceived value of internet, academic and social behaviors. *J. Comput. Hum. Behav.* **82**, 16–33 (2018)



Legal Analysis of the Right to Privacy Protection in the Age of Artificial Intelligence

Sun Xin, Pei Zhaobin^(✉), and Qu Jing

College of Law and Humanities, Dalian Ocean University, Dalian 1160231, Liaoning, China
pzb@dlou.edu.cn

Abstract. The rise of artificial intelligence has had many impacts on the existing society, among which the right to privacy is a social issue worthy of attention. A large amount of personal information, especially personal privacy, is collected from multiple sources, passed through a back-room operation, and then combined, as a new source of value, it also poses a great threat to the security of citizens' right of privacy. Individual citizens face not only this direct link between Internet service providers such as Baidu, but also the fact that behind the web, the individual citizen is also indirectly related to many subjects in the data industry chain, such as the data broker, the data producer, and so on.

This paper is divided into three parts: the first part introduces the overview of the right to privacy; the second part analyzes the challenges of the protection of the right to privacy in the age of artificial intelligence, including the expanding scope of the object of the right to privacy, the severity of the consequences of the infringement, the new changes in the way of infringement, the accountability of infringement is more difficult; The third part puts forward the countermeasures to improve the protection of privacy in the age of artificial intelligence, improve the existing privacy laws and regulations, standardize the legal system, strengthen the industry self-discipline, strengthen international cooperation.

Keywords: Digital age · Privacy right legal protection · Legal countermeasures

1 The Introduction

With the advent of the era of artificial intelligence, the inherent boundary between the field of artificial intelligence and physical space has been broken by digital information and communication technology, the principle of "privacy before the door" and "privacy within the home" are incompatible with the operating model of the AI era, limiting the exclusive control and domination of individuals over the private sphere, the right to privacy overlaps with the protection of personal information, and the protection of the right to privacy in the age of artificial intelligence is faced with many difficulties. The overall planning of data development and utilization [1], privacy protection and public security has become an important guidance and basis for privacy protection in the era of artificial intelligence.

2 Overview of the Right to Privacy in the Age of Artificial Intelligence

2.1 The Meaning and Development of the Right to Privacy

The Origins of the Right to Privacy. The term of privacy has a long history. From the view of Western Christian philosophy, the reason why people know shame is one of the signs that people are different from other things besides having their own thoughts. The earliest origins of privacy are found in the Western Bible, where Adam and Eve went up to a tree to pick leaves and make clothes to wear. The concept of privacy is relative to the individual, we all have the freedom not to be disturbed, disturbed; in our country, privacy is traditionally called “private” secret information, it also refers specifically to sexual relations between men and women. Privacy is the object of the right to privacy, but the right to privacy is a modern term. In 1890, American jurists Warren and Brandeis first used the concept of “privacy” in a joint paper [2] the right to privacy is a unique right: “it means that due to the development of traditional media such as cameras and newspapers, the information about the individual’s right to portrait and the right to health is leaked and disseminated, which disturbs the normal life of some individuals and leads to the demand of the rule of law about the protection of private information.” With the development of society and economy, especially information science and technology, the right of privacy has gradually expanded to the fields of information privacy, space privacy and privacy self-determination [3]. The concept of privacy has been studied for more than 100 years in the fields of philosophy, psychology and sociology. To some extent, privacy is a kind of culture, which is deeply influenced by the culture, morality and economy of a nation, different regions and countries have different ideas about the meaning, scope and protection of privacy.

The Meaning of the Right to Privacy. The object of the right of privacy is privacy, which is clearly written into the legal provisions in our country. The legal guarantee of personal safety and property safety is as old as the common law with the development of social politics and economy, especially the improvement of people’s cultural accomplishment [4], “a world of intangible property formed in the process of spiritual products or thinking has been opened up and recognized by law; Gossiping and the act of breaking the moral bottom line to make public personal affairs should be prohibited,” said Brandeis, Warren and other scholars in the article on the right to privacy, in the modern civilized society, each of us as an individual, privacy is the object of the right to privacy, which is the legal interest protected by the right to privacy, it is still the theoretical basis and basis of the rule of law of our right of privacy, and is one of the essential elements to maintain social order, human dignity and personal independence [5].

2.2 New Changes in Concepts Related to the Right to Privacy in the Age of Artificial

Intelligence Expanding the Scope of Privacy: Inferable Information. At present, artificial intelligence is widely used, and the original regulations on the protection of privacy can no longer meet the needs of social development, relying on the development of big data, artificial Intelligence has a particular role in the processing of information. It can combine and analyze vast, fragmentary, independent, and undirected information, and finally deduce a series of data information, which is purposeful, in most cases, big data can invade a person's privacy [6]. For example, it can predict an individual's spending habits, buying preferences, and even economic level based on the browsing history left behind by the web, companies and other institutions can use this information to push goods in a targeted way that infringes people's right to privacy, at which point the definition of privacy needs to be broadened.

Change in the Nature of the Right to Privacy: Both as a Property Property. Generally speaking, we regard the right to privacy as a right of personality. When most privacy rights are violated, they are often related to moral interests, and the resulting economic losses will be compensated according to the standard of compensation for moral damage. With the continuous development of artificial intelligence, People's right to privacy has also been greatly damaged. For example, as in the case of the 2018 Facebook leak, most businesses sell to other businesses or individuals the personal information they collect from their customers in their business and service activities, in order to achieve the goal of data sharing and cooperation agreement, thereby seeking economic benefits. In the data transaction like this, the user's private information becomes the object of the transaction, and the transaction subject achieves the goal of seeking the property benefit. At this point, the right to privacy has an obvious property attribute.

The Status Quo of the Legal Protection of Privacy in China. Compared with the European and American countries, our country's privacy protection starts late. The Constitution does not directly stipulate that the right to privacy is a fundamental right, only from the perspective of article 39, "inviolability of citizens 'houses' and article 40, 'Citizens' Freedom of communication and privacy protected by law" the right to privacy is protected. Until 2009 "Tort Law of the People's Republic of China" promulgated, the right to privacy by the independent personality protection. At the same time, in the aspect of criminal law, our country also takes the personal information as the protected legitimate rights and interests, and consummates the personal information protection through the criminal law. With the rapid development and popularization of the internet, the state attaches more importance to the protection of personal information. In 2016, the Standing Committee of the National People's Congress passed the Cybersecurity Law of the People's Republic of China. The new general provisions of the General Provisions of the Civil Law of the People's Republic of China enacted by the NPC in 2017 makes up for the shortcomings of the General Principles of the Civil Law of the People's Republic of China, which explicitly defines the right to privacy as an independent civil right to be protected. Article 110 of the Civil Code of the People's Republic of China provides: The personal information of a natural person shall be protected by law. Any organization or individual needing to obtain the personal information of other persons

shall legally obtain and ensure the security of such information, and shall not illegally collect, use, process, or transmit the personal information of other persons, nor illegally buy, sell, provide, or publish the personal information of other persons.

3 Challenges to the Protection of Privacy in the Age of Artificial Intelligence

3.1 Expansion of the Scope of the Object of the Right to Privacy

With the wide application of digital technology, we are under the invisible surveillance all the time. The right to privacy is not only what we must face now, but also a series of unexpected problems with the development of artificial intelligence. The network and the data are the artificial intelligence time interactive main carrier, the privacy right object includes the following several aspects.

The Identity, Property, and Health Status of a Person Logged in. Before the era of artificial intelligence, the object of privacy is mainly personal identity card information, contact, the basic information of family members, family income and other information. In the age of artificial intelligence, when an individual applies for online services such as wechat, QQ, shopping, health care, dating, nailing, Tencent conferences, etc. Service providers often require users to log on to their names, identities, health status and other information; they must provide their bank cards, credit cards, account numbers, passwords and other property information when conducting online financial transactions and making payments, it makes it easier for citizens to divulge their privacy.

Web Activity Tracking. In the age of artificial intelligence, the web has become the biggest source of data. Our habits and activities are captured by Baidu, Google, Kuaishou and volcano videos Our shopping is constantly monitored by Tmall, vipshop, Doddo, jd.com and other major shopping sites; Our social and chat histories are constantly tracked by QQ, E-mail, Weibo and wechat. Flash cookies give away information about our ge habits or location, advertisers then track this information and push targeted sales ads [7].

Daily Routine. Our daily activities are monitored, our smartphones monitor where we are, our workplaces, events, stores, communities and so on. The development of digital sensor technology has made it possible to collect data on a daily basis, such as automated payment systems for parking lots and license plate recognition systems [8], implantable sensors to monitor the health of patients [9], surveillance systems to monitor the elderly at home, and so on, as sensor technology matures, various types of sensors will be widely used in our individuals and organizations [10].

3.2 The Consequences of the Infringement Will Be More Serious

Infinite Zoom and Download. The network has the characteristics of openness and real-time, which makes the information in the network have the effect of infinite amplification. When people publish information on the Internet, it can be immediately disseminated around the world and downloaded countless times. If the artificial intelligence age of personal privacy security line is broken, millions of personal information will be "streaked" in front of internet users around the world.

The “Dividend” of Re-use of Private Information. The value of privacy in the era of artificial intelligence does not lie in the basic use of personality rights, but more in the property of secondary use. The digital trace left by the user will form a complete picture, including interest, orientation, demand and other personality images will be fully displayed, and the advertisers will predict the user's movements, can provide more targeted services, access to huge economic benefits. In recent years, due to the leakage of citizens' personal information caused by counterfeit credit cards, illegal mortgages, malicious marketing caused property, personal injury cases are common, network fraud has become the main type of fraud cases.

The Leakage of Privacy in Cyberspace Often Leads to the Damage of the Right of Reputation in Reality. Take public figures, public figures have a strong influence on society or recognized status in academia or on behalf of the state in the exercise of functions and characteristics, it is easy to cause concern. For example, the “Hollywood porn gate” incident, many popular film star nude photos leaked, leading to the global spread; Hong Kong film and television actor Carina Lau was exposed as a brutal nude photos, triggering strong condemnation from all walks of life in Hong Kong; During the outbreak of Doctor Zhang Wenhong's case was pushed to the forefront of public opinion caused nationwide attention.

3.3 New Changes Have Taken Place in the Pattern of Infringement

Copyright Infringement Is High Tech. In the era of artificial intelligence, the infringement of privacy rights relies on high-tech means. The infringer must have a high degree of professional knowledge of computer network and skilled operation skills, otherwise, it is difficult to break into other people's systems and steal other people's data. The infringement means high-tech, the infringement behavior time is short, the infringement evidence is the invisible information which does not leave the mark. Ordinary users lack of network knowledge, it is difficult to find and collect specific evidence of infringement [11].

Infringement is More Subtle. In the era of artificial intelligence, more and more intelligent devices are installed in personal space, and these devices are equipped with infrared sensors and information sensing devices such as GPS Global Positioning System, and connected to the Internet, open these devices to upload the collected data, and then through the algorithm of data integration and analysis, you can achieve intelligent positioning and monitoring. In addition, the seemingly fragmentary and worthless information collected through analysis and processing, but the information can be processed

and integrated, directly to personal privacy information. This kind of invasion of privacy is usually not easy to find. In addition, in the process of using the Internet, people inadvertently enter phishing sites or Trojan links, steal personal information, or some normally generated non-directional information, after the technology processing of unknown ports, it may be some targeted private information, but it is not clear to the infringer. The development of artificial intelligence has made access to information much different than in the past, and much more insidious than traditional methods of invasion of privacy.

Privacy Data is Vulnerable to Misuse and Proliferation. At present, the personal privacy data is not only controlled by the government in the process of performing the public management service function, but also controlled by many internet companies. These companies misuse, sell or share large amounts of private data for profit. Once these data leaks have spread, it is difficult to trace back to how many subjects have access to them. It is difficult to get effective relief after the invasion of personal privacy.

3.4 Accountability for Violations is More Difficult

It's Expensive and Difficult. Under the background of "everything is connected", various ports are everywhere. So when privacy is compromised, it's hard to know which port is compromised. Even with the consent of the party concerned, the party consciously provides and allows the other party to legally obtain general personal information, on a specific port, using artificial intelligence for data analysis, and then synthesize it into personal information with a specific sexual orientation. But it is hard and expensive for the aggrieved party to prove it.

The Tortfeasor Is Difficult to Identify. In general, there are three ways in which AI can invade privacy. First, the infringers consciously use artificial intelligence to invade personal privacy and achieve their improper purposes. At this point, artificial intelligence as a tool or means used by the infringer in the process of infringement, the infringer can be directly investigated for legal responsibility. Second, AI is flawed in design, manufacturing and marketing, leading to invasion of privacy. In such cases, the provisions on product liability may apply to the duty of investigation. Third, "artificial intelligence" goes beyond the individual infringement of human control. At this point, it is necessary to study whether artificial intelligence has the qualification of infringement subject, whether it can bear the tort liability independently.

The Damage is Serious and Difficult to Repair. In the age of artificial intelligence, if individual privacy is divulged and diffused, it is very difficult to remedy completely. Personal privacy information is held by several ports, which port, how much general, directional information, when and where to use the technical power to integrate into a specific personal privacy information, is hard to know for sure. Therefore, in the era of artificial intelligence, privacy has been violated, generally cannot be restored to the initial state of fullness, can not be reversed. In addition, taking advantage of the risk of privacy violations in cloud computing, cloud computing can provide a shared pool based on demand use of resources, so whether individuals, businesses, or governments,

can store large amounts of data in the cloud. Stored in the cloud, private information is at risk of being stolen and leaked, and is vulnerable to potential threats such as hacker attacks.

4 Under the Artificial Intelligence Age to the Privacy Protection Legal Countermeasure Research

4.1 Improving Existing Laws and Regulations on the Right to Privacy

China's current law on the right to privacy lag, lack of norms to protect the public's right to privacy in the era of artificial intelligence. First of all, as the Basic Law of our country, the Constitution of the People's Republic of China to privacy in the constitution mainly focuses on articles 38, 39 and 40, but there is no provision on the right to privacy in artificial intelligence. Therefore, the Constitution should set up a privacy protection clause under AI. However, due to the abstractness and principle of the constitution, it is difficult to regulate the privacy of artificial intelligence in detail. Secondly, the artificial intelligence privacy right is brought into the scope of civil law protection, that is, the protection of the artificial intelligence privacy right is defined as a special form of privacy right. In addition, the criminal law should distinguish the right of privacy and the right of artificial intelligence privacy in detail in order to perfect the existing laws and regulations and protect the right of privacy only by protecting the information of citizens. Article 10 of the Civil Code of the People's Republic of China does not adequately protect the right to privacy of artificial intelligence. The privacy right of artificial intelligence can be protected as an independent personality right, its legal status should be clarified, and a complete protection system should be established. Finally, China should enact separate legislation on artificial intelligence privacy right as soon as possible, such as "the Development Plan on the New Generation of Artificial Intelligence", in order to change the past privacy protection lag, scattered, fuzzy, one-sided plight, make laws and regulations better protect related rights and interests. In addition, after the establishment of a more complete legal framework, according to the characteristics of the artificial intelligence era, the elements of the right to privacy of artificial intelligence should be carefully regulated and supplemented, and the establishment conforms to the actual damage many kinds of relief way and the way, causes the laws and regulations to be closer to the reality.

4.2 The Legal System Needs to Be Improved

Step up Regulation. In order to follow the trend of the development of artificial intelligence, the state council issued the "Development Plan for the new generation of artificial intelligence" in 2017, in order to seize the development opportunities, regulate and guide the development of artificial intelligence. At present, China has taken legislative measures to protect personal information, but in the era of artificial intelligence, China lacks a unified legal document that specifically regulates the protection of personal information, in order to more flexibly adapt to the threat and impact of personal privacy information,

more strictly adapt to the healthy development of artificial intelligence, and effectively regulate and guide. In the process of formulating relevant laws, we should further clearly define the scope of protection of individual's right to privacy, and enrich the strength of individual's management and control of information, regulate and restrict the storage and utilization of private information and data collected by enterprises or governments. Clear artificial intelligence in the design and manufacture process of the rights and obligations of different subjects, in the application of artificial intelligence, the infringement of the right to privacy of the responsibility of the subject made clear provisions. While improving the legal provisions, we should strengthen the supervision of the enterprises, the government and other subjects who hold personal data to obtain and use these data illegal acts. We will establish a special regulatory agency for the whole process of AI design and development, data collection and application, form an open and transparent regulatory system, and actively bring in relevant experts to strengthen effective supervision in the event and after the event. Minimize the consequences of invasion of privacy. Some countries and regions in the world have begun to take corresponding legislative measures to deal with the difficulties and challenges brought by the legal system of protecting individual privacy. For example, the EU's Common Data Protection Regulation, which came into force in 2018 and was introduced in 2016, explicitly grants individuals rights over data, such as the right to decide how they are to be used and treated. Therefore, countries should combine their basic national conditions, as well as the stage of artificial intelligence development, absorb and learn from the international legal norms on the protection of the right to privacy more advanced countries.

Creation of Punitive Liability Clauses. The cases of data leakage and infringement of privacy rights involve not only the protection of data private interests, but also the maintenance of data public interests. For the consequences caused by mass infringement of data public interests, compensation liability cannot be used to compensate and recover, it is necessary to defend the privacy rights of non-specific subjects in digital space with punitive damages, increase the punitive clauses of unlimited maximum amount, and strike down severely the punitive power of violating the privacy rights in the artificial intelligence age. At the same time, in the process of dealing with cases of data breach and privacy infringement, it is not conducive to the recovery of damages suffered by non-specific subjects by promoting punishment alone, and through the compatible application of punishment for data violation and the conciliation system, to form a three-dimensional accountability system for privacy infringement.

4.3 Strengthening Self-discipline in the Industry

Strengthening Government Administrative Supervision. To improve the protection of network privacy in the era of artificial intelligence, the common function of the protection of network privacy and legislation is the self-discipline of the industry, which plays an indispensable auxiliary role, only the combination of the two can go hand in hand, can give full play to the protection network privacy right maximum effect. The scope of the industry self-discipline convention should not only be confined to the network daily behavior norms and the network privacy protection, but also strengthen

the government's administrative supervision of the industry self-discipline. We should give full play to its own functions of guidance and standardization, set up a special supervisory organization, endow the supervisory organization with exclusive authority, make the supervisory procedure transparent, and supervise the self-discipline of the network industry effectively.

Give Full Play to the Technological Advantages of the Industry. In view of the inferior position of government administrative affairs organs in terms of technology and talents, trade associations have just grasped the development problems and current situation of the industry, and can take on the role of a bridge between the government and Industry Enterprises, provide professional advice to the government in formulating laws, regulations and administrative policies in a timely manner.

4.4 Strengthening International Cooperation

The challenge of the rapid development of artificial intelligence to the right of privacy is a problem that all countries in the world are facing. Therefore, countries should also strengthen exchanges and cooperation on the common problems in the development of artificial intelligence, such as security risk issues, ethical issues, etc., can form general security risk standards and technical standards, unify the norms and guide the design and development of artificial intelligence, in order to prevent the chain reaction of crisis between countries under the influence of economic globalization, we should carry out the global management of artificial intelligence, promote the healthy development of artificial intelligence and better serve human society.

5 Conclusion

The right of privacy is formed in the traditional society, which is characterized by the right of privacy, the right of domination and the right of exclusiveness. With the development of technology, artificial intelligence has been rapid development, although human life has a lot of convenience, but the individual privacy of citizens also brought a huge impact and challenge. However, the protection of the right to privacy is not achieved overnight, especially with the development of information technology, the boundaries and scope of privacy are constantly changing, and the protection of the right to privacy is a dynamic process, on the protection of the right to privacy legal developments to follow up in a timely manner, to explore the improvement. Therefore, it is necessary for individuals, the public, the government and even the whole world to make concerted efforts to deal with the challenges brought to the human society by the artificial intelligence era and reduce the adverse impact of artificial intelligence, to better serve the development of human society.

References

1. Resolution on the outline of the 14th Five-Year Plan for economic and social development and long - range objectives through the year 2035 (2021). http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm. Accessed 29 May 2022

2. Warren, S.D., Brandeis, L.D.: Right to privacy. *Harv. L. Rev.* **4**(5), 193–220 (1890)
3. Wang, L.: Re-definition of right to privacy. *Jurist* **24**(1), 108–120, 178 (2012)
4. Brandeis, L.D., Warren, S.D.: *The Right to Privacy*. Peking University Press, Beijing (2014)
5. Shen, Z., Xu, W.: *On the Right to Privacy and Personality*. Shanghai People's Publishing House, Shanghai (2010)
6. Wang, J.: The dilemma and the way out of privacy protection in the age of artificial intelligence. *North Media Res.* **3**(4), 46–50 (2018)
7. Soltani, A., Canty, S., Mayo, Q., Thomas, L., Hoofnagle, C.J.: Flash cookies and privacy. In: *Proceedings of the AAAI Spring Symposium on Embedded Reasoning: Intelligence in Embedded Systems*, pp. 158–163. AAAI, Palo Alto (2010)
8. Foresti, G.L., Mähönen, P., Regazzoni, C.S.: *Multimedia Video-Based Surveillance Systems: Requirements Issues and Solutions*. Springer, Berlin (2000)
9. Maheu, M., Allen, A., Whitten, P.: *E-Health, Telehealth, and Telemedicine: a Guide to Startup and Success*. John Wiley & Sons, San Francisco (2001)
10. Beckwith, R.: Designing for ubiquity: the perception of privacy. *IEEE Perv. Comput.* **2**(2), 40–46 (2003)
11. Zhang, L.: Protection of privacy from the perspective of civil code-focus on challenges and solutions of privacy in the era of big data. *West. L. Rev.* **32**(2), 88–98 (2021)



An Intrusion Detection Method Fused Deep Learning and Fuzzy Neural Network for Smart Home

Xiangdong Hu^(✉), Qin Zhang, Xi Yang, and Liu Yang

College of Industrial Internet of Things, Chongqing University of Posts and Telecommunications, Chongqing, China
huxd@cqupt.edu.cn

Abstract. Smart home depending on Internet of things (IoT) technologies is facing severe risks in information security. An intrusion detection method fused deep learning and fuzzy neural network for smart home is proposed, according to the method, the data features are built by deep learning methods, the high-dimensional data is mapped into low-dimensional one, and the categories of attack can be analyzed and distinguished based on fuzzy neural network. A method used in optimizing network depth also is provided, and this can overcome the problem from the traditional method which determines the layer number of network depending on experience. The simulation results show that the proposed method including artificial intelligence can improve the detection rate of attacks, for example, the detection rate can reach 94% for the denial of service attack and remote illegal access, and the detection rate of the tested new types of attacks in the network exceeds 60%.

Keywords: Smart home · Intrusion detection · Deep learning · Fuzzy neural network

1 Introduction

With the rising of “Digital China” and “Smart City”, smart home is an example of deep integration between information technology and home environments. However, while people enjoy various conveniences brought by smart home, they are plagued by multiple information security threats resulted from smart home, such threats including illegally accessing the smart home network and stealing user privacy data, implementing denial of service attacks by taking advantage of vulnerabilities of Internet of Things equipment, maliciously controlling network, or injecting illegal instructions, and so on. These attacks will cause the smart home network to operate improperly, even to threat the user’s body or property safety. Therefore, it is necessary to timely discover any attacks in a smart home network and some appropriate countermeasures must be taken. As an active defense means, intrusion detection is an important method to guarantee network security [1].

In recent years, some scholars at home and abroad have done multi-angle researches on security issues in smart home, mainly including: Chen proposed a smart home security

monitoring strategy based on GPRS, and its research is limited to the realization of alarm mechanism for smart home by controlling hardware [2]. Wu proposed a secure transmission model based on the trusted computing technology and the signature-encryption method of bilinear pairings [3]. Nobakht proposed an intrusion detection and mitigation framework (called IoT-IDM) to provide a network-level protection for smart devices deployed in home environment [4]. It can detect the attacked hosts and correspondingly generate responses on identifying any attack source by machine learning and a linear logistic regression classification model. The strategy is applied to the perception layer to avoid further attacks on the IoT devices. Although this solution can achieve a high detection rate, but it only works well for certain specific application scenarios. Umer proposed an intrusion detection technology based on network flow, and he analyzed packet header characteristics and the statistical characteristics of packets through principal component analysis and time-series statistics to determine whether attacks occur on the network [5]. This technique targeting the obvious flow characteristics behaviors has a better detection effect, the permission bypass and cross-site attack detection is poor.

This paper proposes a multi-layer neural network intrusion detection method based on deep learning and fuzzy neural network. It seeks to achieve high detection rate and low false detection rate based on the new artificial intelligence method.

2 Intrusion Detection Model

2.1 Network Composition of Smart Home

The smart home based on the Internet of Things technology uses various home devices. The sensor nodes in the home device can complete the transmission of measurement and control instructions and information exchange in home device. The number of sensor nodes involved in a smart home is generally small, and the node locations are relatively fixed, the network of smart home network generally constitutes as shown in the Fig. 1.

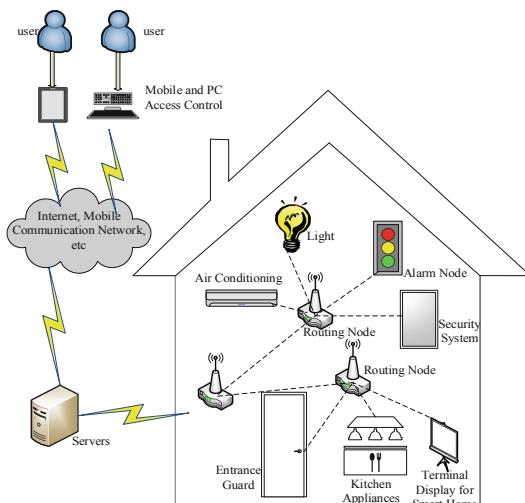


Fig. 1. The structure of smart home network

2.2 Intrusion Detection Algorithm

The core of intrusion detection model for smart home is an intrusion detection algorithm composed by multi-layer neural networks. The multi-layer neural network structure shown in Fig. 2, and the network is composed of a plurality of the Restricted Boltzmann Machines and a fuzzy neural network. The core steps include: the learning of Restricted Boltzmann Machines to update parameters; the determining the network depth; the adjustment of multi-layer neural network weights.

2.3 The Restricted Boltzmann Machines

The Restricted Boltzmann Machine [6] is a two-layer neural network whose two layers of nodes respectively are visible layer nodes and hidden layer nodes. The structure of the Restricted Boltzmann Machine is shown in Fig. 2. As the name shown, the connection is limited, so there are no connections between every sides, then there is only a side connection between the visible units and the hidden units. And V is the visible layer unit, indicating input data; H is a hidden layer unit, indicating no practical meaning but it can generate by machine learning automatically; W indicates the connection weight between the visible layer unit and the hidden layer unit; A indicates the visible layer offset; B indicates hidden Layer offset. When V is input, the hidden layer H is obtained through $P(H|V)$. After the hidden layer H is obtained, the visible layer can be obtained again through $P(V|H)$. If the reconstructed visible layer V is the same as the original V by adjusting the parameters, the hidden layer obtained is another representation of the visible layer.

Each node in a Restricted Boltzmann Machines has only two states, namely {0, 1}, where the state of the node is 1, indicating that the node is activated; the state of the node is 0, indicating that the node is not activated. The state of each layer node is determined by comparing it with a random number. So if the probability value of node activation is greater than a randomly generated number, it is considered that the node is activated, and the value is 1; conversely, the node is not activated and the value is 0.

Because nodes in the same layer are mutually independent, all hidden layer nodes are also mutually independent with visible layer known, so and the probability distribution of the j-th node in the hidden layer can be expressed as:

$$\begin{cases} P(H|V) = \sum_j p(h_j|v) \\ p(h_j = 1|v) = f(b_j + \sum_i p(h_j|v)) \\ p(h_j = 0|v) = 1 - p(h_j = 1|v) \end{cases} \quad (1)$$

Similarly, with the hidden layer known, all visible layer nodes are conditionally independent, so the probability distribution of the i-th node in the visible layer can be expressed as:

$$\begin{cases} P(V|H) = \sum_i p(v_i|h) \\ p(v_j = 1|h) = f(c_j + \sum_j w_{ij} h_j) \\ p(v_j = 0|h) = 1 - p(v_j = 1|h) \end{cases} \quad (2)$$

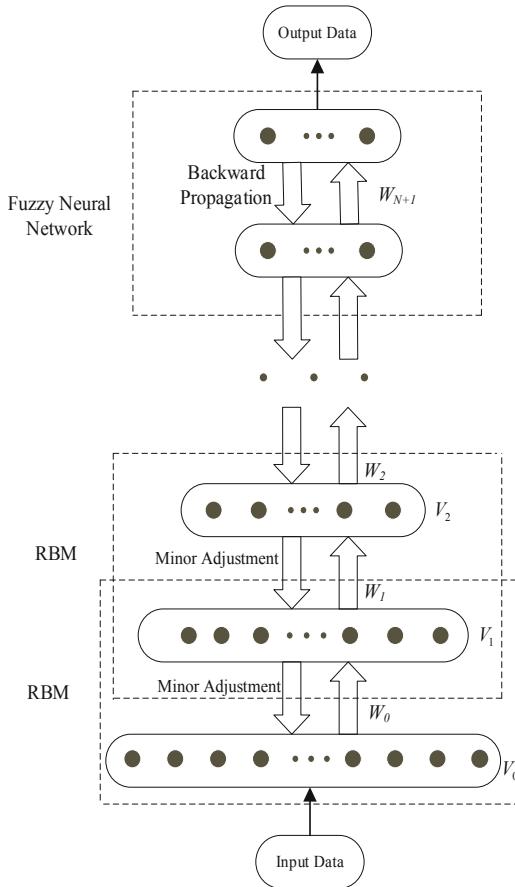


Fig. 2. The structure of multilayer neural network

where the function f is an activation function (sigmoid):

$$f(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

2.4 The Fast Learning Algorithm Based on Contrast Divergence

The task of Restricted Boltzmann Machines learning is to find the optimal W , A , B , and the distribution obtained by the Restricted Boltzmann Machines network fits input data as best possible.

The specific algorithm for limiting the Boltzmann machine to k-step Gibbs sampling is as follows: initialize the visible node state $v^{(0)}$ with a training sample, alternating the following sampling:

$$\begin{aligned} h^{(0)} &\sim P(h|v^{(0)}) \\ v^1 &\sim P(v|h^{(k-0)}) \\ &\dots \dots \\ v^{(k)} &\sim P(v|h^{(k-1)}) \\ h^{(k)} &\sim P(h|v^{(k)}) \end{aligned} \tag{4}$$

Using $P(v|h^{(k-1)})$, the sample is $v^{(k)}$

Using $P(h|v^{(k)})$, the sample is $h^{(k)}$

The approximate estimate of Eq. (4) using v_k obtained by k-step Gibbs sampling can be approximately written as:

$$\begin{aligned} \Delta w_{ij} &= \frac{\partial \ln p(v)}{\partial w_{ij}} \approx p(h_i = 1|v^{(0)})v_j^{(0)} - p(h_i = 1|v^{(k)})v_j^{(k)} \\ \Delta a_j &= \frac{\partial \ln p(v)}{\partial a_j} \approx v_j^{(0)} - v_j^{(k)} \\ \Delta b_i &= \frac{\partial \ln p(v)}{\partial b_i} \approx p(h_i = 1|v^{(0)}) - p(h_i = 1|v^{(k)}) \end{aligned} \tag{5}$$

The method to update the weights is:

$$\theta_{N+1} = \theta_N + \mu \frac{\partial \ln p(v)}{\partial \theta} \tag{6}$$

where N represents the number of iterations and μ is the learning rate.

2.5 Determination of Depth Value of Multi-layer Neural Network

In this paper, the depth of multi-layer neural network is determined by referring to the concept of reconstruction error, and training data is used as the initial state, then once the Gibbs sampled difference quantity is used to define the reconstruction error through the Restricted Boltzmann Machines distribution.

$$R_{cerror} = \frac{\sum_{i=1}^n \sum_{j=1}^m (v_{ij} - v_{ij}^{(1)})}{nm} \tag{7}$$

The specific rules are:

$$\begin{cases} L_{deep} = N_{RBM} + 1, R_{cerror} > \in \\ L_{deep} = N_{RBM}, R_{cerror} < \in \end{cases} \tag{8}$$

where ϵ is the threshold value of the reconstruction error, and L_{deep} is the number of hidden layers. If the network passes the training at this time, the reconstruction error is lower than the threshold, then the reverse adjustment of the parameters is started; otherwise, the network depth is increased by 1 and the training is continued. The reconstruction error judgment method is simple and easy to implement.

2.6 Weight Minor Adjustment of Multi-layer Neural Network

The T-S fuzzy neural network is defined by the “if-then” rules [7]. In the case of the rule R^i , the fuzzy inference is as follows:

$$\begin{aligned} R^i \text{ if } x_1 &\text{ is } A_1^i, \dots, x_k \text{ is } A_k^i \\ \text{then } y_i &= p_0^i + p_1^i x_1 + \dots + p_k^i x_k \end{aligned} \quad (9)$$

where A_j^i is a fuzzy set, $p_j^i (j = 1, 2, \dots, k)$ is a fuzzy neural network parameter, y_i is an output obtained from a fuzzy neural network. The input part is fuzzy, but the output part is determined, and the fuzzy inference represents the output as a linear combination of inputs.

The T-S fuzzy neural network algorithm is a supervised classifier that uses the output error to evaluate the error of the immediate previous layer of the output layer. This error is used to evaluate the more recent error, passing through the layer by layer to learn, then get error estimates for other layers.

Multi-layer neural network weight minor adjustment process:

Step 1. Calculate the actual output y' of the output node for v_i in each training sample;

Step 2. Calculate the error gradient $\delta_k = y'(1 - y')(y - y')$ between the actual output of the output node and the ideal output (y);

Step 3. Calculate the error gradient of the hidden layer unit h

$$\delta_h = y'(1 - y') \sum_k \theta_{hk} \delta_k \quad (10)$$

where θ_{hk} is the connection weight value from node h to the subsequent node k;

step 4. Calculate weight updates

$$\begin{aligned} \theta_{ij} &= \theta_{ij} + \Delta\theta_{ij} \\ \Delta\theta_{ij} &= \mu O_i \delta_j \end{aligned} \quad (11)$$

where μ is the learning rate, determined by experiment and experience; O_i is the output of node i, and δ_j is the recursive error gradient of node j.

3 Simulation and Experiment

Using the Matlab platform and JavaWeb to construct the smart home network intrusion detection model, determined the depth of the multi-layer network and evaluated the function and performance of the multi-layer neural network.

3.1 Data Set Preprocessing

The simulation data came from KDDCUP99 intrusion detection data set of MIT Lincoln Laboratory, the data set was divided into training data and test data, the data set label types were divided into Normal, Dos, Probing, R2L, and U2R. The bold label indicated that it only exists in the test set, as shown in Table 1.

Table 1. The tag type of data set

Type	Label content	Meaning
Normal	normal	Normal data
Dos	smurf, teardrop, pod, apache2, processstable	Denial of service attack
Probing	ipsweep, nmap, portsweep, satan, mscan,saint	Port scanning or monitoring
R2L	ftp_write, imap, guess_passwd, phf, spy, multihop, warezclient, warezmaster, xlock, worm, named	Illegal access by remote machines
U2R	buffer_overflow, loadmodule, perl, rootkit, xterm, ps, http tunnel	User's illegal access to local superuser privileges

According to the proportion of various tags in the original data set, 111950 records were randomly selected from the training data set to generate a training set for layer neural network training, in which Normal, Dos, Probing, R2L, and U2R were 30000, 80000, 1000, 800, and 150 records, respectively.

In accordance with the proportion of various types of tags in the original data set, randomly selected 81400 records from the test data set, of which Normal, Dos, Probing, R2L, and U2R were 20000, 60000, 500, 800, and 100 records, respectively. Performing the following processing for each record: mapping character types to numeric types; the data is normalized and the size of the data is reduced to the range [0,1]; the recorded labels are coded as 0 for Normal, 1 for DoS, 2 for Probing, 3 for R2L, and 4 for U2R.

3.2 Experiments and Analysis

We need to determine the depth of the multi-layer neural network, which is the number of Restricted Boltzmann Machines that are used, and then use the test set to evaluate the performance of the intrusion detection. This paper will use the detection rate p_d of attack behavior and the false alarm rate p_f of normal behavior. The performance of the algorithm is evaluated in terms of the accuracy of all behavior detection p_e and the time required to identify all the data in the test set. Its definition is as follows:

$$p_d = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

$$p_f = \frac{FP}{FP + TN} \times 100\% \quad (13)$$

$$p_e = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (14)$$

where TP means true positive (actual attack data are classified as attacks) and FP are false positives (actual normal data are classified as attacks). Likewise, FN stands for false negative (actual attack data are classified as normal) and TN means true negative (normal data are classified as normal).

1) Depth Determination of Multi-layer Neural Networks

The training set was selected to use the multilayer neural network proposed for the depth of 2, 3, 4, and 5, using 1, 2, 3, and 4 Restricted Boltzmann Machines for training multilayer neural networks respectively. The test set is used as the input to the trained multi-layer neural network to implement the mapping of the original 41 data into 10-dimensional data. Then the fuzzy neural network was used to classify low-dimensional data obtained by multilayer neural networks of different depths. Training set reconstruction error, test set accuracy and test set running time, the results were shown in Table 2.

Table 2. The performance indicators of multi-layer network with different depths

Depth	Reconstruction error	Accuracy/%	Runtime/s
2	29.63	85.2	5
3	8.93×10^{-2}	91.8	8
4	1.06×10^{-2}	94.3	12
5	5.64×10^{-3}	92.6	20

At the beginning of the experiment, the accuracy rate increased as the depth deepens. When the network depth increased to a certain value, the detection rate decreased. With the increase of the network depth, the detection time would increase, the time complexity and space complexity would increase. The reconstruction error threshold was 0.02 in this paper. From the above Table, considering the detection accuracy and detection time, we can see that when the network depth was equal to 4 to meet the conditions. Combined with smart home server performance, a multi-layer neural network with a depth of 4 is chosen in this paper.

2) Intrusion Detection Performance Analysis

The neural network proposed in this paper was a multi-layer neural network composed of a limited number of the Restricted Boltzmann Machines and a fuzzy neural

network. Under the premise of ensuring its detection rate, it could detect new types of attacks in the network. The multi-layer neural network proposed in this paper was compared with shallow neural network BP neural network and literature [8] based on deep learning intrusion detection system detection rate and false detection rate. The results were shown in Fig. 3 and Fig. 4.

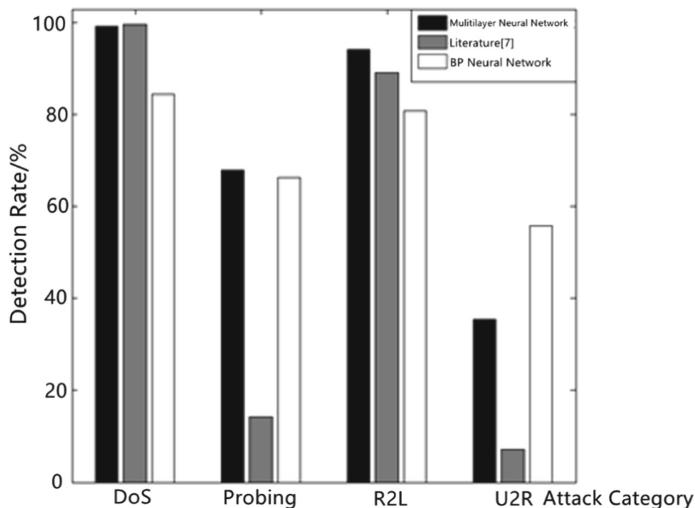


Fig. 3. The comparison of detection rate

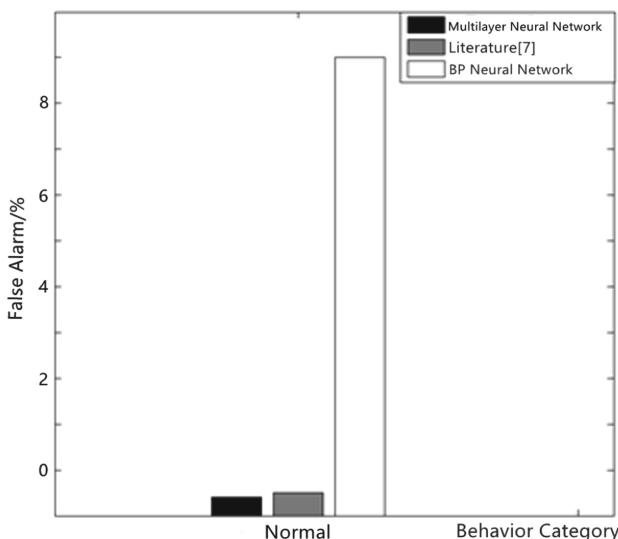


Fig. 4. The comparison of false alarm rate

By Fig. 3, the multi-layer neural network and the deep learning layer 4 neural network mentioned in [8] had higher detection rate and lower false positive rate for large sample size DoS and R2L attacks, and the detection rate of the multi-layer network is best. The detection rate of probing attack and U2R attack with a small sample size was lower, the detection effect of shallow network was better than that of deep network. However, because the multi-layer neural network proposed in this paper introduced a fuzzy neural network, the deep learning method was improved. The ability of small samples, and the detection rate was higher than the method in [8]. 100, 500, 1000, 5000, 10000, 50000, and 100000 test set data were selected according to the proportion of various types of attacks to train the multilayer neural network. The obtained test set data detection accuracy rate is shown in Fig. 5.

Combining with the characteristics of smart home networks, the attacker attacked the network in the physical detection platform to form attacks that had not appeared in the network, imitate an attacker to crack the user's login password through a brute force cracking tool, the DDoS attacks such as SYN_FLOOD and UDP_FLOOD were initiated to the server through related software. The smart home server gateway data packet was captured and the data set was generated statistically. A new type of attack in which the KDDCUP99 data set did not appear in both the training set and the test set was selected and 100 test sets were formed. The test set was applied to the multilayer neural network proposed in this paper, a deep belief network composed of four Restricted Boltzmann Machines, and a BP neural network. Their detection rates are 62%, 38% and 28%, respectively.

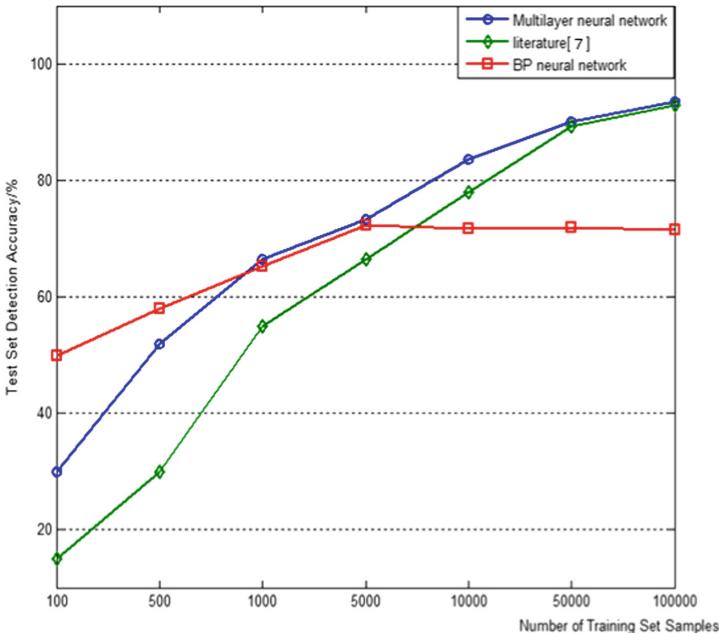


Fig. 5. The contrast between detection accuracy and numbers of training samples

4 Conclusion

Aiming at the technical characteristics of current smart home, a smart home intrusion detection algorithm based on multi-layer neural network is proposed by integrating multiple Restricted Boltzmann Machines and fuzzy neural networks. Simulation and physical test results show that the proposed method has a higher detection rate than the existing methods, the detection rate for DoS and R2L is over 94%, and the false detection rate for normal behavior is less than 1% and with a detection rate of more than 60% for the new type of attack tested, this method demonstrates the performance advantages and good adaptability of the smart home network in intrusion detection.

References

1. Yang, Y., Huang, H., Shen, Q., et al.: Research on intrusion detection based on incremental GHSOM. *Chin J. Comput.* **37**(5), 1216–1224 (2014)
2. Chen, S., Zhong, X., Liu, J., et al.: Safety monitoring for intelligent living-room based on GPRS. *Comput. Measur. Control* **19**(2), 326–328 (2011)
3. Wu, Z., Zhou, Y., Ma, J.: A security transmission model for Internet of Things. *Chin. J. Comput.* **34**(8), 1351–1364 (2011)
4. Nobakht, M., Sivaraman, V., Boreli, R.: A host-based intrusion detection and mitigation framework for smart home IoT using open flow. In: 2016 11th International Conference on Availability, Reliability and Security (ARES), pp. 147–156. IEEE (2016)
5. Umer, M.F., Sher, M., Bi, Y.: Flow-based intrusion detection: technique sand challenges. *Comput. Secur.* **70**, 238–254 (2017)
6. Hinton, G.E., Sejnowski, T.J.: Learning and relearning in Boltzmann machines. *Parallel Distril. Process.* **1**, 282–317 (1986)
7. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern.* **1**, 116–132 (1985)
8. Alrawashdeh, K., Purdy, C.: Toward an online anomaly intrusion detection system based on deep learning. In: 2016 15th IEEE International Conference on Machine Learning and Applications, pp. 195–200 (2016)



A High Performance Intrusion Detection System Using LightGBM Based on Oversampling and Undersampling

Hao Zhang^{1,2}, Lina Ge^{1,2(✉)}, and Zhe Wang^{1,2,3}

¹ School of Artificial Intelligence, Guangxi Minzu University, Nanning, China
66436539@qq.com

² Key Laboratory of Network Communication Engineering, Guangxi Minzu University, Nanning, China

³ Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning, China

Abstract. Intrusion detection system plays an important role in network security, however, the problem with data imbalance limits the detection ability of intrusion detection system. In order to improve the performance of intrusion detection system, this paper proposes to use the adaptive synthetic sampling technique (ADASYN) and random under sampling technique to alleviate the problem of data imbalance in intrusion detection. Firstly, the majority class samples in the dataset are removed by undersampling technology and the minority class samples are oversampled, so the samples can reach a balanced state. Subsequently, a sparse autoencoder (SAE) extracts features from the resampled data to fit the original sample as closely as possible. Finally, LightGBM is applied on the processed dataset for the classification process. Multi-classification experiments were conducted on KDD99 and UNSWNB15 datasets. We compare six models' performance and find LightGBM is superior to other models. Furthermore, we also compare existing methods and the results show that our proposed method outperforms current methods.

Keywords: Intrusion detection systems · Resampling · LightGBM · Autoencoder

1 Introduction

With the advent of the big data era, network traffic is on an exponential growth trend. The growing data makes it increasingly difficult to detect intrusion traffic. As a security device, the intrusion detection system plays a vital role in protecting the country's information infrastructure [1]. According to the detection technology, intrusion detection systems can be divided into misuse-based intrusion detection systems (MIDS) and anomaly-based intrusion detection systems (AIDS). The former detects anomaly traffic based on historical traffic statistics and known attack libraries, while the latter detects anomaly traffic based on statistical deviations [2, 3]. MIDS is unable to detect unknown attacks and is vulnerable when dealing with zero-day attacks. AIDS can detect unknown

attacks, so it has become a popular research topic. In recent years, machine learning in intrusion detection has achieved success [4]. Typical methods such as decision tree (DT) [5], support vector machine (SVM) [6], random forest (RF) [7], etc., have shown high accuracy in the detection of abnormal traffic. Compared with traditional methods, machine learning technology can efficiently detect attacks in the face of massive traffic data, so it exerts an important influence in intrusion detection. As an extension of machine learning techniques, deep learning has gained widespread attention owing to its ability to extract features [8–11]. It can learn low-dimensional expressions from high-dimensional data so that massive data can be processed. The popular methods are deep belief networks (DBN) [12, 13], Autoencoders (AE) [14, 15], which are efficient in feature extraction.

In the era of big data, the issue that needs to be addressed in intrusion detection is how to identify the attack traffic with only contains a small number of traffic. Generally, in the network flow, most of them are normal traffic, and only a few are anomaly traffic. It is also in line with people's observations in life. The currently available public datasets for intrusion detection include binary and multiclass. Most of these datasets do not have imbalance problems on binary classification. However, different types of attacks require different defense mechanisms [16], and simply dividing traffic into normal and abnormal traffic cannot allow managers to take targeted defense measures. Therefore, it is necessary to identify the traffic to a specific type. Although machine learning technology can be handy in coping with massive data, it is inefficient in the detection of minority classes.

An efficient way is to use resampling techniques [17, 18]. In the multi-classification task, only a few data are attack classes, which makes the classifier tend to the majority classes when training data. Although it has high performance in general, it is especially low when detecting the minority classes. Resampling technology can address this problem. For the minority classes, the oversampling technology is used to increase the number of the minority class data. At the same time, the undersampling technology is used to reduce the number of majority class data. After data resampling, the number of majority class and minority class samples reaches a balanced state, so that the classifier can detect minority class samples. However, how to ensure that the resampling samples come from the same data distribution as the real samples is a key issue of resampling techniques. In general, the resampling technology adopts the Euclidean distance to judge whether the data come from the same distribution. The method assumes that spatially close samples are from the same distribution. Whereas the use of Euclidean distance ignores the outliers, making the resampling technique have errors in fitting the real samples. Deep learning techniques provide a solution for this problem. Using deep learning technology for feature extraction can alleviate the problems caused by data fitting.

There have been many studies on data imbalance [16, 19–23]. Particularly, in this paper, an intrusion detection method based on data imbalance is presented. The method adopts ADASYN oversampling and random undersampling techniques. The ADASYN technique oversamples the minority class samples and can help the classifier to learn a better decision boundary. In addition, a random undersampling technique is used in this paper to remove the samples from most classes. A combination of oversampling and undersampling methods enables the classifier to detect minority class samples. After oversampling and undersampling, a sparse autoencoder is designed to extract features.

The autoencoder is trained from raw data, so it is helpful for data fitting problems caused by data resampling. The LightGBM is used as the classifier. LightGBM is a boosting ensemble method, which is improved based on GBDT. LightGBM gives more weight to the misclassified samples so that they can get more attention in the weak classifiers later. For minority class samples, LightGBM can improve the detection rate. Histogram and feature bundling methods were introduced into the LightGBM, which can reduce the training time while obtaining high accuracy. Therefore, the LightGBM model is used for classification in this paper. The main contributions of this study are as follows:

- (1) To solve the data imbalance problem, we proposed a hybrid sampling method to process the dataset and used the LightGBM for classification. This method alleviates the imbalance problem.
- (2) We adopted sparse autoencoders to perform feature extraction. In addition, we designed a resampling method to make the resampled data closer to the actual distribution. It is shown in Sect. 3.2.
- (3) We adopted several metrics, including accuracy rate, precision rate, detection rate, and F1 score to evaluate the performance of the method. Furthermore, the proposed method was compared not only with classical methods but also with current advanced methods.

The rest of the paper is organized as follows: Sect. 2 introduces the theory related to ADASYN techniques and LightGBM. Section 3 presents the proposed method in detail. Section 4 is the experimental part of the paper, which describes and analyzes the experimental results. The final part is the conclusion and outlook.

2 Related Works

2.1 Adasyn

The ADASYN method oversamples minority classes on the boundary to move the decision boundary to the minority classes that are difficult to learn [24]. The specific steps of the method are as follows: Assume that the training set D_{tr} has m samples $\{x_i, y_i\}^m, y_i \in \{+1, -1\}$. There are m_s minority class samples and m_l majority class samples in the training set D_{tr} , $m_s + m_l = D_{tr}$.

- (1) Calculate the imbalance ratio in the D_{tr} :

$$d = \frac{m_s}{m_l} \quad (1)$$

- (2) If $d < d_{th}$ (d_{th} is the preset threshold), enter the loop:

- a. Calculate the number of minority class samples that need to be generated ($\beta \in [0, 1]$, is a balance parameter):

$$G = (m_l - m_s) \times \beta \quad (2)$$

- b. For each minority class sample, K nearest neighbor samples are picked based on the euclidean distance (Δ_i represents the number of samples that K nearest neighbor samples belong to the majority class)

$$r_i = \frac{\Delta_i}{K} (i = 1, 2, \dots, m_s) \quad (3)$$

- c. Calculate the weights:

$$\hat{r}_i = r_i / \sum_i^{m_s} r_i \quad (4)$$

- (3) Calculate the number of samples that need to be synthesized for each minority class sample:

$$g_i = \hat{r}_i \times G \quad (5)$$

- (4) For each minority class sample x_i , randomly select a sample x_{zi} from its K nearest neighbor samples to generate a new sample according to the following formula:

$$s_i = x_i + (x_{zi} - x_i) \times \lambda (\lambda \in [0, 1]) \quad (6)$$

2.2 LightGBM

Among the tree-based ensemble algorithms, GBDT, XGBoosting, and LightGBM are recognized as the best performing algorithms. In recent years, they have made a big splash in Kaggle competitions. GBDT and XGBoost algorithms need to traverse the entire dataset when dividing nodes, which makes them have a large time overhead. Compared with other algorithms, LightGBM has a greater advantage in training overhead. LightGBM adopts gradient-based one-sided sampling (GOSS) method and exclusive feature bundling (EFB) algorithm [25], making it significantly smaller in training time than GBDT and XGBoosting.

Assuming that each sample is associated with its gradient information, it is generally believed that a sample with a small gradient has a low training error. For the following classification or regression tasks, samples with small gradients have smaller contribution values. Therefore, the GOSS method sorts the gradients of all samples according to their absolute value and saves all samples with large gradients. Subsequently, randomly selects some samples from small gradient samples for preservation. Suppose there are n training instances in the training set, denoted as $\{x_1, \dots, x_n\}$. After the gradient boosting calculation, the negative gradient of the output of the model is recorded as $\{g_1, \dots, g_n\}$. The information gain of the split feature j at point d is defined as:

$$V_{j|O}(d) = \frac{1}{n_O} \left(\frac{\left(\sum_{\{x_i \in O: x_{ij} \leq d\}} \right)^2}{n_{l|O}^j(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} \right)^2}{n_{r|O}^j(d)} \right) \quad (7)$$

where O is the parent node of the decision tree division, n_l is the left node of the decision tree, and n_r is the right node of the decision tree, and $n_O = \sum I[x_i \in O]$, $n_{l|O}^j(d) = \sum I[x_i \in O : x_{ij} \leq d]$ and $n_{r|O}^j(d) = \sum I[x_i \in O : x_{ij} > d]$.

Let a and b be the sampling ratios of large gradient and small gradient instances, respectively. According to the sorted instance gradient values, the first $a \times 100\%$ large gradient sample is selected, and then randomly selects $b \times 100\%$ small gradient samples from the rest of the data. After many iterations, the final calculated information gain is:

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i)^2}{n_l^j(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i)^2}{n_r^j(d)} \right) \quad (8)$$

where $A_l = \{x_i \in A : x_{ij} \leq d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \leq d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$.

EFB method exploits the sparsity of high-dimensional spaces to reduce features. In high-dimensional space, the values between two instances are mostly mutually exclusive, that is, they are not all zero in the same row. Therefore, an undirected graph with weights can be further constructed. The nodes of the graph represent a feature, and the weights are expressed as the number of mutually exclusive values between two features. It is then transformed into a graph coloring problem and solved using a greedy algorithm.

3 Methodology

Network flow presents the characteristics of massive and high dimensions. In order to improve the detection ability, an intrusion detection method based on oversampling and undersampling is designed in this paper. The schematic diagram of the method is shown in Fig. 1. The proposed method includes data preprocessing module, resampling module, feature extraction module, and classification module.

In the data preprocessing module, these methods including numeralization, one-hot coding, and normalization are used for data processing. After data preprocessing, it is partitioned into a training set and test set. It is described in Sect. 3.1.

In the resampling module, this paper uses the ADASYN and random undersampling technology to make the training data reach a relatively balanced state. This is described in detail in Sect. 3.2.

In the feature extraction module, a sparse autoencoder is trained from the original data. Then the autoencoder is used to feature extraction on the resampled data, which reduces the data dimension while decreasing the fitting error caused by data resampling. The details of feature extraction are presented in Sect. 3.3.

In the classification module, the LightGBM model is used to train the data after processing so that the trained LightGBM model is obtained. Finally, the test set is input into the LightGBM model for testing, and the final results are output.

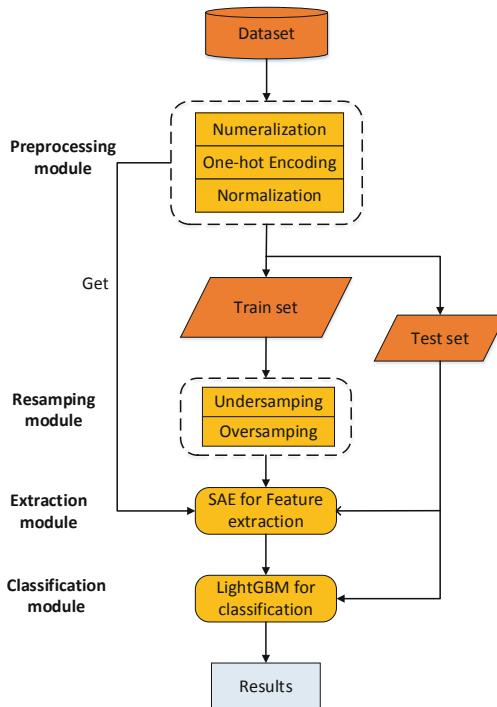


Fig. 1. A schematic diagram of the intrusion detection system

3.1 Dataset Description and Preprocessing

The KDD99 dataset comes from the intrusion detection project at Lincoln laboratory in the United States. A total of nine weeks of network connection data was collected in the dataset, which contained 7 million records. The dataset contains a total of 39 attack types, of which 22 appear in the training set and 17 in the test set. These 39 attack types can be classified into four categories, namely Probe, Dos, R2L, and U2R. 10% of the KDD99 dataset is used in this paper. On the basis of this dataset, it is further divided into training set and test set with a ratio of 7:3.

The UNSWNB15 dataset was developed by the Australian Cyber Security Centre in 2015 using the IXIA tool. The dataset contains 2 million records, which are saved in four CSV files. The dataset contains a total of 9 attack types. In this paper, all the data in this dataset are used for experiments. Similarly, the UNSWNB15 dataset is also divided into training set and test set in a ratio of 7:3.

The KDD99 dataset contains 41-dimensional features, of which the three features of protocol, service, and flag are symbol types, and the others are numeric types. Since the classifier cannot process symbolic data, the symbolic data is first converted into numeric types. Then the data are sparsely encoded by one-hot encoding. Taking the KDD99 dataset as an example, the protocol feature contains three values: TCP, UDP, and ICMP, which are encoded as 100, 010, and 001 respectively. The service feature has 66 values, so it is represented using 66 numbers consisting of 0 and 1. The flag feature has 11

values, which are represented by 11 numbers containing 0 and 1. After one-hot coding, the dimensionality of the KDD99 dataset is extended to 118. Similarly, the UNSWNB15 dataset is processed in the same way.

Dataset normalization processing. To speed up the training, the data normalization method is required. In this study, the maximum and minimum normalization method is used to scale the data to [0,1]. The maximum-minimum normalization method is calculated as follows:

$$\text{MaxMinScaler} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (9)$$

Among them, x_{\max} and x_{\min} represent the maximum and minimum values of the column where the feature x is located, respectively.

3.2 Resampling Method

In order to obtain a proper representation of the data distribution in practice, this paper does not simply resample all data to the same amount. Instead, the majority and minority class data are resampled to a relatively balanced state. First, a threshold θ is preset, which represents an acceptable balance ratio. In general, the threshold is set to 0.5. For a dataset, its balanced degree is determined by the following equation:

$$d = \frac{|T_i|}{|T|} \quad (10)$$

where $|T_i|$ represents the number of instances of the i -th type of data, and $|T|$ represents the total number of samples.

For each class of samples, its balance degree is calculated according to the formula (10). If the balance degree exceeds the threshold θ , the samples are undersampled. The sample size is calculated as:

$$|T'_i| = \frac{|T_i|}{C} \quad (11)$$

where C represents the number of categories in the dataset, and $|T'_i|$ is the number of samples of this category after undersampling.

After that, sort all $|T'_i|$ and save the smallest $|t'_i| = \text{Min}(|T'_i|)$. For the minority class whose balance degree is less than θ , it needs to be oversampled. The sample size for oversampling is calculated as:

$$|T'_i| = |T_i| + 10^N \quad (12)$$

where N is the number of digits of $|t'_i|$.

After the above calculations, the number of samples required for each class is obtained. To ensure that the number of samples for all classes is in the same order of magnitude, it needs to be traversed. For classes that are not in the same order of magnitude, they are further processed so that they are in the same order of magnitude as other samples.

3.3 Feature Extraction and Classification

After data resampling, a sparse autoencoder is used to extract features from the data. First, in the data preprocessing module, training a sparse autoencoder so that the autoencoder can extract the low-dimensional representation of the data. The trained sparse autoencoder will be used to recover a low-dimensional representation of the resampled data. The structure of the sparse autoencoder used in this paper is shown in Fig. 2. A sparse autoencoder consists of an input layer, an encoding layer, a middle layer, a decoding layer, and an output layer. The size of the input and output layers is determined by the dimensionality of the input data. Both the encoding and decoding layers are composed of two neural network layers with sizes of 80 and 50, respectively. The middle layer consists of neurons with size 16 and is a low-dimensional representation of original data. After feature extraction, the data will be input into the LightGBM classifier training. Finally, the test data will be input into the LightGBM model for testing.

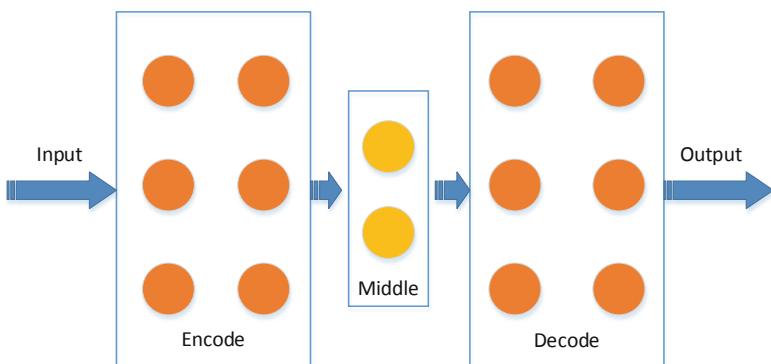


Fig. 2. The structure of the sparse autoencoder

4 Experiment and Discussion

This part presents the performance of the LightGBM model before and after the data resampling. Furthermore, to validate the performance of LightGBM, the paper compared with support vector machine, random forest, GBDT, and so on. The experiments were deployed on a Dell host with 32G memory and were programmed using Python 3.8. The sklearn library and the keras framework were used to build the model.

4.1 Evaluate Metrics

In this paper, the four metrics of accuracy rate (AC), precision rate (P), detection rate (DR), and F1 scores (F1) are used to evaluate the model.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$P = \frac{TP}{TP + FP} \quad (14)$$

$$DR = \frac{TP}{TP + FN} \quad (15)$$

$$F1 = 2 / \left(\frac{1}{P} + \frac{1}{DR} \right) \quad (16)$$

4.2 Results and Analysis

We conducted multi-classification experiments on KDD99 and UNSWNB15 datasets. The experiments compared the LightGBM's performance before and after data resampling. To ensure the comparability of the methods, the random_state was set to 42. Note that all results were rounded to two decimal places.

The performance of LightGBM is shown in Fig. 3. As shown in Fig. 3, the accuracy, precision, and F1-score of the model on the KDD99 dataset are improved by 0.46% after data resampling. In addition, the accuracy, precision, and F1-score of the model on the UNSWNB15 dataset are improved by 0.63%, 1.13%, and 0.83% respectively. It can be concluded that the performance of the model has improved after data resampling.

Figure 4 and Fig. 5 show the detection rates of each class on the two datasets before and after data resampling. It can be seen from Fig. 4 that the detection rates of Normal and Dos classes are above 90% before resampling because of their high proportion in the original dataset. The detection rates of Probe and R2L classes are lower than 80% before resampling. After resampling, the detection rate increased significantly. In particular, the detection rate of U2R before resampling is zero. The reason is that the number of U2R in the original data is rare, with only 52 records. After resampling, the detection rate of U2R reaches 94.12%. The performance of the model on the UNSWNB15 dataset is shown in Fig. 5. After resampling, the detection rate of Exploits decreases, indicating that the boundary between this data and adjacent samples becomes blurry. However, the model improves the detection rate to varying levels in the remaining attack classes. In general, the detection of minority classes is improved after resampling.

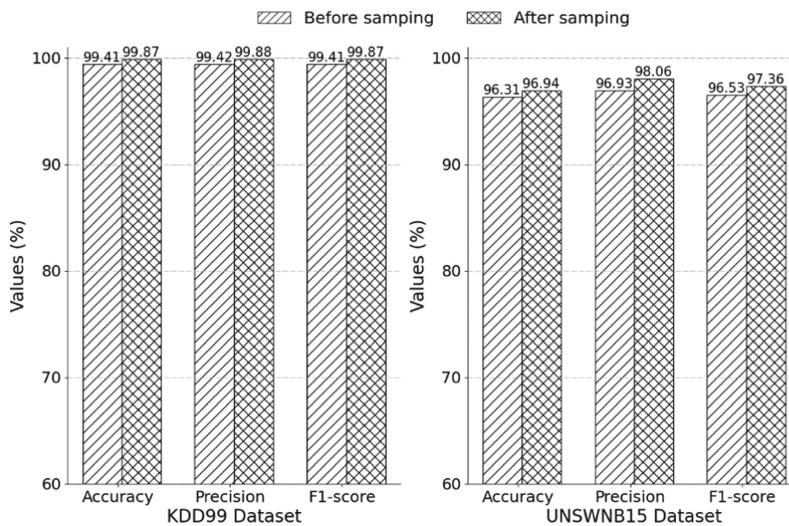


Fig. 3. The performance of model before and after resampling for KDD99 and UNSWNB15 datasets

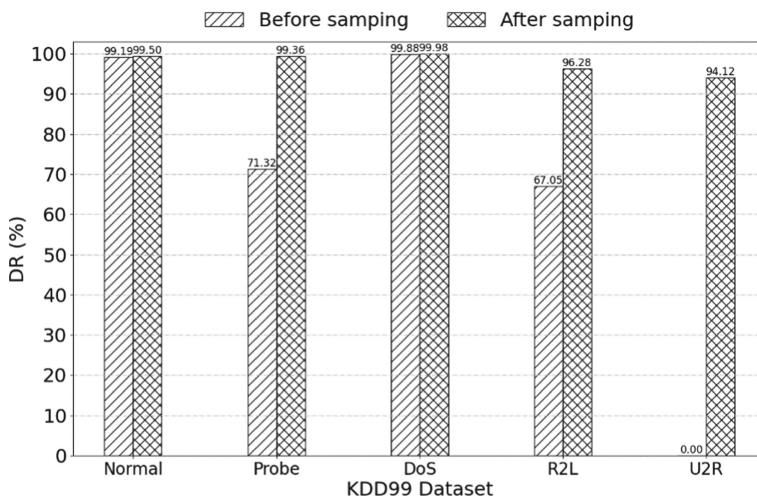


Fig. 4. The detection rates of each class for the KDD99 dataset

Tables 1 and 2 show the performance of different models after resampling. On the KDD99 dataset, the LightGBM model achieves 99.87% accuracy, which is the highest among all models. In addition, the model obtains 99.88% and 99.87% in precision and F1 scores, respectively. The RF model also performs the same in these metrics. However, the training time for the LightGBM is significantly less than for the RF model. Though the DT model has the lowest training time among all models, its accuracy, precision, and F1 scores are lower than the LightGBM model. On the UNSWNB15 dataset, although

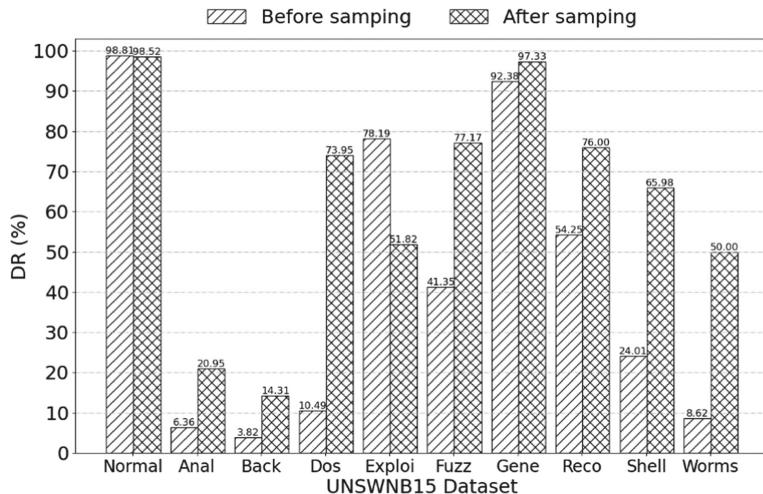


Fig. 5. The detection rates of each class for the UNSWNB15 dataset

the RF and XGBoost models have an outstanding performance in some metrics, the training time of RF is 6 times longer than the LightGBM model and the training time of XGBoost is 7 times longer than the LightGBM model. In general, the LightGBM model balances time overhead and overall performance on both the KDD99 dataset and the UNSWNB15 dataset.

Table 1. The performance of different models on the KDD99 dataset after resampling

Model	SVM	DT	RF	GBDT	XGBoost	LightGBM
AC (%)	99.31	99.66	99.86	99.31	99.84	99.87
P (%)	99.69	99.71	99.88	99.60	99.86	99.88
F1 (%)	99.46	99.68	99.87	99.42	99.85	99.87
Time (s)	72.66	1.03	15.75	156.96	13.97	1.56

Table 2. The performance of different models on the UNSWNB15 datasets after resampling

Model	SVM	DT	RF	GBDT	XGBoost	LightGBM
AC (%)	96.62	96.64	96.99	96.73	96.95	96.94
P (%)	97.86	97.73	97.92	97.92	98.02	98.06
F1 (%)	97.07	97.10	97.36	97.19	97.36	97.36
Time (s)	8412.03	3.97	65.42	735.57	76.41	10.09

Tables 3 and 4 show the detection rate for specific classes. On the KDD99 dataset, all models have identical performance in the detection of the Dos and Probe classes. On the detection of Normal, LightGBM and RF models perform the best. On the detection of R2L, XGBoost and LightGBM have the highest detection rate. GBDT and SVM perform best in the detection of U2R. Overall, the LightGBM model outperforms in all classes, although detection on U2R is not the best. On the UNSWNB15 dataset, for classes of Analysis and Backdoors, all models have a low detection rate, no more than 50%. The possible reason is that the boundaries of Analysis and Backdoor samples are relatively blurred, which makes the model unable to distinguish them effectively. The detection rates of LightGBM models are above 50% for all classes excluding Analysis and Backdoor classes. However, the SVM and GBDT models do not reach 50% in the detection of Exploits. DT, RF, and XGBoost models also do not reach 50% in detecting Worm classes. It means that the LightGBM model is more effective than other models in detecting minority classes.

Table 3. The detection rates of different models for each class on the KDD99 dataset after resampling

Model	SVM	DT	RF	GBDT	XGBoost	LightGBM
Normal	0.97	0.99	1.00	0.97	0.99	1.00
Probe	0.99	0.98	0.99	0.99	0.99	0.99
Dos	1.00	1.00	1.00	1.00	1.00	1.00
R2L	0.85	0.92	0.95	0.90	0.96	0.96
U2R	1.00	0.76	0.88	1.00	0.94	0.94

Table 4. The detection rates of different models for each class on the UNSWNB15 dataset after resampling

Model	SVM	DT	RF	GBDT	XGBoost	LightGBM
Normal	0.98	0.98	0.99	0.98	0.99	0.99
Dos	0.76	<u>0.55</u>	0.51	0.74	0.70	0.74
Exploits	<u>0.43</u>	0.54	0.61	<u>0.49</u>	0.54	0.52
Fuzzers	0.67	0.71	0.79	0.71	0.78	0.77
Generic	0.97	0.97	0.97	0.97	0.97	0.97
Backdoors	0.16	0.04	0.04	0.12	0.15	0.14
Reconnaissance	0.77	0.63	0.75	0.72	0.76	0.76
Analysis	0.18	0.16	0.19	0.17	0.20	0.21
Worms	0.66	<u>0.24</u>	0.26	<u>0.59</u>	<u>0.40</u>	0.50
Shellcode	0.45	0.52	0.63	0.66	0.67	0.66

Tables 5 and 6 show the detection rates compared to other advanced methods. For the KDD99 dataset, it is clear from Table 10 that FE-SVM has the poor performance. Among all the methods, our proposed method has the best performance in detecting the U2R class. The RU-Smote method performs best in detecting the R2L class. Overall, our proposed method achieves an average detection rate of 97.8% and is more effective on the KDD99 dataset compared to other methods. For the UNSWNB15 dataset, the TSDL, CASCADE-ANN, and DBN methods achieve a 0% detection rate on some classes. By contrast, ICVAE-DNN, SAE, and our proposed method can detect all classes. In particular, our proposed method achieves an average detection rate of 62.6%, which is the highest among all methods. This proves the generalizability of our proposed method.

Table 5. Comparison multi-class classification results with advanced methods on the KDD99 dataset.

Method	FE-SVM[6]	ACAE-RF[7]	PDAE[15]	RU-Smote[18]	MSML[23]	Our method
Normal	0.99	1.00	1.00	0.98	0.86	1.00
Probe	0.70	0.99	0.99	0.99	0.98	0.99
Dos	0.96	1.00	1.00	1.00	1.00	1.00
R2L	0.45	0.88	0.92	0.97	0.91	0.96
U2R	0.14	0.47	0.75	0.83	0.73	0.94
Avg	0.648	0.868	0.932	0.954	0.896	0.978

Table 6. Comparison multi-class classification results with advanced methods on the UNSWNB15 dataset.

Method	TSDL[8]	CASCADE-ANN[10]	ICVAE-DNN[11]	DBN[13]	SAE[14]	Our method
Normal	0.82	0.80	0.81	0.70	0.78	0.99
Dos	<u>0.00</u>	<u>0.00</u>	0.08	<u>0.00</u>	0.06	0.74
Exploits	0.57	0.57	0.71	0.86	0.94	0.52
Fuzzers	0.40	<u>0.00</u>	0.35	0.57	0.52	0.77
Generic	0.61	0.98	0.96	0.96	0.97	0.97
Backdoors	<u>0.00</u>	<u>0.00</u>	0.21	<u>0.00</u>	0.04	0.14
Reconnaissance	0.25	0.34	0.80	0.73	0.81	0.76
Analysis	0.01	<u>0.00</u>	0.15	<u>0.00</u>	0.01	0.21
Worms	<u>0.00</u>	<u>0.00</u>	0.80	<u>0.00</u>	0.11	0.50
Shellcode	0.01	<u>0.00</u>	0.92	<u>0.00</u>	0.59	0.66
Avg	0.267	0.269	0.579	0.382	0.482	0.626

5 Conclusions

For the problem that the abnormal traffic in the network flows is less than the normal traffic, we propose adaptive synthetic oversampling technology and random undersampling technology to process imbalanced data. In the experiments with two datasets, the model performance before and after data resampling was compared. After resampling, the LightGBM achieves 99.87% accuracy, 99.88% precision, and 99.87% F1 score on the KDD99 dataset, respectively. On the UNSWNB15 dataset, it achieves 96.94%, 98.06%, and 97.36% for accuracy, precision, and F1 scores respectively. The overall performance of LightGBM is improved after resampling. Meanwhile, the paper also compares the performance of different models after data resampling. The results show that our proposed method not only performs better but also has a low time overhead. Furthermore, the paper compares the proposed methods with state-of-the-art methods, and the results are promising. Although the LightGBM model performs well in the overall performance, it still needs to be improved in detecting minority classes such as Analysis and Backdoors. In the future, we will further optimize the model so that it can exceed 50% detection rate for these classes.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant 61862007, and Guangxi Natural Science Foundation under Grant 2020GXNSFBA297103.

References

1. Bijone, M.: A survey on secure network: intrusion detection & prevention approaches. *Am. J. Inf. Sys.* **4**(3), 69–88 (2016)
2. Jian, S.J., et al.: Overview of network intrusion detection technology. *J. Inf. Secur. China* **5**(4), 96–122 (2020)
3. Mahmoud Said, E., et al.: A novel hybrid model for intrusion detection systems in SDNs based on CNN and a new regularization technique. *J. Netw. Compu. Appl.* **191**, 103160 (2021)
4. Kilincer, I.F., et al.: Machine learning methods for cyber security intrusion detection: datasets and comparative study. *Comput. Netw.* **188**, 107840 (2021)
5. Ahmim, A., et al.: A novel hierarchical intrusion detection system based on decision tree and rules-based models. In: 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 228–233. IEEE (2019)
6. Dong, K., Shi, J., Guo, L., Yuan, F.: Application of SVM in anomaly detection based on sampling and feature extraction. *J. Phys. Conf. Ser.* **1629**(1), 012017, IOP Publishing (2020)
7. Ji, S., Ye, K., Xu, C.-Z.: A network intrusion detection approach based on asymmetric convolutional autoencoder. In: Zhang, Qi., Wang, Y., Zhang, L.-J. (eds.) CLOUD 2020. LNCS, vol. 12403, pp. 126–140. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59635-4_9
8. Khan, F.A., et al.: A novel two-stage deep learning model for efficient network intrusion detection. *IEEE Access* **7**, 30373–30385 (2019)
9. Kasongo, S.M., Sun, Y.: A deep learning method with wrapper based feature extraction for wireless intrusion detection system. *Comput. Secur.* **92**, 101752 (2020)
10. Baig, M.M., Awais, M.M., El-Alfy, E.M.: A multiclass cascade of artificial neural network for network intrusion detection. *J. Intell. Fuzzy Syst.* **32**(4), 2875–2883 (2017)

11. Yang, Y., et al.: Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network. *Sensors* **19**(11), 2528 (2019)
12. Zhao, Z., Ge, L., Zhang, G.: A novel DBN-LSSVM ensemble method for intrusion detection system. In: 2021 9th International Conference on Communications and Broadband Networking, pp. 101–107 (2021)
13. Singh, P., Kaur, A., Aujla, G.S., Batth, R.S., Kanhere, S.: Daas: Dew computing as a service for intelligent intrusion detection in edge-of-things ecosystem. *IEEE Internet Things J.* **8**(16), 12569–12577 (2020)
14. Li, Y., Gao, P., Wu, Z.: Intrusion detection method based on sparse autoencoder. In: 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), pp. 63–68, IEEE (2021)
15. Basati, A., Faghih, M.M.: PDAE: Efficient network intrusion detection in IoT using parallel deep auto-encoders. *Inf. Sci.* **598**, 57–74 (2022)
16. Zhang, H., et al.: An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. *Comput. Netw.* **177**, 107315 (2020)
17. Ahsan, R.: A detailed analysis of the multi-class classification problem in network intrusion detection using resampling techniques. Doctoral Dissertation, Carleton University (2021)
18. Bagui, S., Li, K.: Resampling imbalanced data for network intrusion detection datasets. *J. Big Data* **8**(1), 1–41 (2021). <https://doi.org/10.1186/s40537-020-00390-x>
19. Liu, J., Gao, Y., Hu, F.: A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. *Comput. Secur.* **106**, 102289 (2021)
20. Andresini, G., et al.: GAN augmentation to deal with imbalance in imaging-based intrusion detection. *Fut. Gene. Comput. Syst.* **123**, 108–127 (2021)
21. Alshamy, R., Ghurab, M., Othman, S., Alshami, F.: Intrusion detection model for imbalanced dataset using SMOTE and random forest algorithm. In: Abdullah, N., Manickam, S., Anbar, M. (eds.) ACeS 2021. CCIS, vol. 1487, pp. 361–378. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-8059-5_22
22. Gonzalez-Cuautle, D., et al.: Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets. *Appl. Sci.* **10**(3), 794 (2020)
23. Yao, H., et al.: MSML: a novel multilevel semi-supervised machine learning framework for intrusion detection system. *IEEE Internet of Things J.* **6**(2), 1949–1959 (2018)
24. He, H., et al.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328 (2008)
25. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154 (2017)



Research on the Current Situation and Improvement Countermeasures of Farmers' Information Security Literacy Based on New Media

Haiyu Wang^{1,2(✉)}

¹ School of Journalism and Communication, Zhengzhou University, Zhengzhou, China
wanghaiyu0830@163.com

² School of Economics and Management, Zhengzhou Normal University, Zhengzhou, China

Abstract. Research on the current situation and Improvement Countermeasures of farmers' information security literacy under the background of new media information security has become the cornerstone of national overall security in the information age. Improving farmers' information security literacy is an inevitable requirement to protect farmers' personal privacy and property security. Using the questionnaire survey to analyze the four aspects of farmers' information security awareness, information security knowledge, information security ability and information ethics in five provinces, it is found that farmers' information security literacy is not high as a whole, and put forward suggestions and paths to improve information security literacy.

Keywords: Information security · Information security capability · Ethics and morality · Short video

1 Introduction

With the full completion of network infrastructure and the rapid popularization of new media forms and media interaction platforms, the user population of new media has gradually sunk to rural areas. New media is profoundly changing the way farmers live, study and entertainment. According to the data of the 49th statistical report on the development of China's Internet (hereinafter referred to as the report), the scale of rural Internet users in China has reached 284 million, and the Internet penetration rate in rural areas is 57.6%. Farmers and other professional groups share the achievements of informatization development, and can independently complete network activities such as showing health code/travel card, purchasing daily necessities and searching information.

However, on the one hand, farmers enjoy the convenience of communication and information access brought by the rapid development and popularization of new media, as well as the increase of economic income in the form of e-commerce. On the other hand, due to the relative backwardness of rural areas and the limited cultural level of farmers, it has also brought more serious security problems, such as repeated prohibition

of Internet rumors, frequent leakage of personal privacy and even organizational secrets, increasingly serious cyber attacks, and rising cyber crime. Farmers have suffered great economic losses and security threats in these cases, which has become an important factor affecting the sustainable development of rural areas.

In fact, the risks brought by the openness and security vulnerabilities of new media are everywhere. In recent years, news reports about farmers' groups suffering from network fraud, network attack and data leakage have often appeared in the newspapers. However, the research on Farmers' network information security literacy is not rich. Through extensive questionnaire survey, this study takes farmers, a large-scale new media user group, as the research object, peeps into their awareness and ability of new media information security, and puts forward the new media path to improve information security literacy.

The remainder of this paper is organized as follows: Sect. 2 reviews the relevant theories and research review of information security literacy. Section 3 explains the research methods and analyzes the research data. The fourth part is the research conclusions and suggestions.

2 Discussion on Theory and Background

The concept of information security literacy comes from information literacy. Different scholars have made diversified descriptions of information literacy. Chinese scholars have conducted rich research on Farmers' information literacy.

2.1 Information Security Literacy and Division

The concept of information literacy was first proposed by American scholar Paul zurkowski in 1974. The American Library Association accurately stated its definition in 1989, that is, people with information literacy can judge when they need information and know how to obtain, evaluate and effectively use the information they need. In September 2003, the United States Library and Information Commission (NCLIS) and the National Information Forum organized the United Nations information literacy expert meeting and issued the "Prague declaration: towards an information literacy society", pointing out that information literacy has become an important factor in society. How to make people benefit from information and communication resources and technologies in the Internet era is an important challenge facing today's society. American sociologist Ingles put forward that "a country is a modern country only when its people are modern people."

The definition of information security literacy comes from the concept of information literacy. It refers to people's understanding of information security and various comprehensive abilities of information security under the condition of informatization, including information security awareness, information security knowledge, information security ability, information ethics and so on. Among them, information security awareness, knowledge and ethics mainly belong to the cognitive level, while information security ability mainly belongs to the behavioral level. In the context of new media, farmers' information security literacy refers to the information security literacy in the widely

used environment of new media, which mainly involves the information security awareness, information security knowledge, information security ability, information ethics, etc. expressed by individual users in the process of information acquisition, information dissemination and shopping through smart phones and diversified APPs. The author learned in the process of research that some short video users inadvertently revealed their city and unit information in addition to showing the amount of salary in the salary slip during the live broadcast, which caused a lot of trouble for themselves. Another user told the researcher that when she was chatting with her grandmother, her grandmother accidentally picked up her clothes, which was considered by the network supervision to be too revealing, and the user was banned as a result.

2.2 Research Review

According some scholars have studied the relationship between information security and rural social governance. Information security is an important part of network social governance. In rural society, people's awareness of personal information security is not strong, and information related to personal privacy and property interests is not valued. Information distortion is also a thorny problem in the governance of network society. In particular, farmers' ability to identify the authenticity of information is not high, and they are easy to be cheated for the information they see on the Internet, which may lead to property losses. Therefore, in the Internet era, it is necessary to cultivate farmers' information identification ability. The content of network social governance also includes the participation of disorderly networks, including network rumors, network group events and network criminal activities, which will bring new impact on Rural Governance under the background of "Internet + party construction". From the perspective of data security and targeted poverty alleviation, other scholars proposed that improving the security of poverty alleviation data is an important guarantee to enhance the construction of big data poverty alleviation. The "big data + targeted poverty alleviation" model has brought opportunities to China's poverty alleviation. On the other hand, the security of big data has also become an obstacle to poverty alleviation. Therefore, while developing targeted poverty alleviation, we should pay attention to the privacy and data security of poverty alleviation target information in the big data environment, and put forward government legislation, inspection, maintenance and upgrading of information poverty alleviation network, as well as measures to improve data security, such as strengthening farmers' information security publicity and education.

Farmers' information security literacy has received extensive attention from the academic community. Starting from the main body of farmers, this study obtains a wide range of samples through questionnaire survey and interview, analyzes the data by SPSS, so as to spy on the current situation of farmers' information security literacy in the new media environment, and puts forward countermeasures to improve information security literacy, so as to meet the needs of the era of digital village construction.

3 Research Methods and Data Analysis

3.1 Research Methods

Therefore, after training the investigators by recruiting investigators, a questionnaire survey on information security literacy was conducted for farmers in Henan, Shandong and Shanxi provinces. Analyze information security knowledge, information security awareness, information security ability and information security ethics. At the same time, carry out group interviews and in-depth interviews on Farmers' information security. The sampling method scientifically determines the samples by the combination of proportional sampling, equidistant sampling and pure random sampling.

A total of 1000 questionnaires were distributed and 858 were recovered, with a recovery rate of 85.5% 8%. There were 830 valid questionnaires, and the effective rate was 83%. The survey time is from January 2021 to March 2021. Figure 1 shows the proportion of users with education, of which illiteracy accounts for 1.45%; 19% had received primary education only; 53% of them have junior high school education; 20% of them have education in senior high school or technical secondary school or above; Those with high school education or above accounted for 7%. On the whole, the education level of the respondents was generally not high.

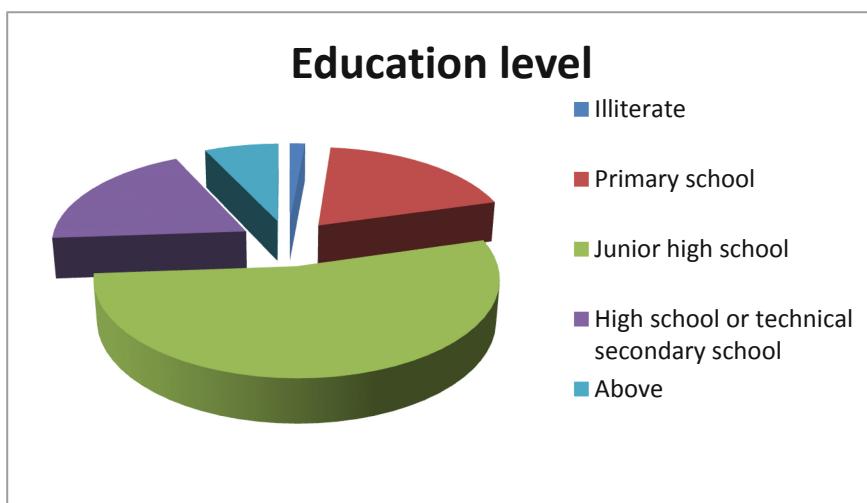


Fig. 1. Education level of farmers

3.2 Investigation and Analysis of Farmers' Information Literacy

(1) Information Security Awareness

Table 1 shows the perception of farmers of different genders on the risk of personal information disclosure in the use of new media and the importance of security software.

The results show that there are significant differences between gender and information security awareness of new media users. There are significant differences between men and women in understanding the permissions of installed applications and paying attention to policies and regulations on information security. Compared with men, women do not know enough about the permissions of installed applications, and pay far less attention to the laws and regulations on information security than men. Men's awareness of information security is higher than women's.

Table 1. Information security awareness of new media users of different genders

Problem	Male	Female	P
Personal information is at risk of being leaked when using new media	92.5%	97.3%	0.079
Apps in the official app store are secure	57.2%	65.4%	0.338
Security software is important	83.6%	83.8%	0.571
Understand the permissions of the installed application	57.6%	36.7%	0.000
Pay attention to information security policies and regulations	49.3%	28.4%	0.002

Table 2. Mastery of information security knowledge of new media users of different genders (%)

Gender	1	2	3	4	5	6	7	8	P 值
Male	43.2	66.4	24	28	24.8	9.6	7.2	16	0.815
Female	49.1	67.2	25	38.8	34.8	13.8	9.4	18.9	

Table 3. Application usage behavior of new media users of different genders (%)

Problem	Gender	Always	Sometimes	Never	P-value
Download and use applications from unknown sources	Male	18.4%	60.8%	20.8%	0.006
	Female	5.71%	66.38%	28.45%	
Grant the requested permission to the application	Male	43.2%	51.2%	5.6%	0.065
	Female	29.31%	62.93%	7.76%	
Click unknown link	Male	14.4%	58.4%	27.2%	0.000
	Female	2.59%	44.83	52.59%	

(2) Information security knowledge

Table 2 shows that there is no significant difference between gender and the mastery of information security knowledge. The overall mastery of information security knowledge

Table 4. Information protection behavior of new media users of different genders (%)

Gender	1	2	3	4	5	6	P-vale
Male	62.4	52.8	55.2	44	47.2	22.4	0.789
Female	68.1	50	71.6	55.6	51.7	18.9	

is poor and needs to be improved. Among them, 1–8 represent the topic of information security knowledge mastery. For example, when 1 is online shopping, personal confidential information (such as bank card number and payment password) is encrypted and transmitted; 2. Antivirus software needs to be installed on the mobile phone, otherwise it is easy to be infected with virus when accessing the network; 3 chat records on social software cannot be obtained by others except the parties.

(3) Information security capability

There are significant differences between gender and information security behavior of new media users. Table 3 shows that women are better off than men in downloading and using apps from unknown sources with only 5.71% of women often doing so, compared with 18.4% of men. On the issue of clicking on unknown links, the proportion of men who often click is also higher than that of women, 14.4% and 2.59% respectively, while the proportion of men who never click is much lower than that of women, 27.2% and 52.59% respectively. The difference is very significant. There is no significant difference in the permissions requested by authorized applications. There is no significant difference in information protection behavior. The above data shows that although men have stronger information security awareness than women, their performance in information security behavior needs to be strengthened, which also verifies the research results of S. Allam. Awareness does not necessarily translate into actual behavior.

1–6 are the items to measure the user's information protection behavior, for example, 1 is to install security software on the mobile phone; 2. Data backup; 3 turn off location services, etc.

(4) Information security ethics

The survey on the attitude towards public legal constraints on the Internet shows that 44% of the respondents do not know that all their speech acts on the Internet are under the supervision of the law; 33% of the respondents knew it, but it was not serious; Only 24% of the respondents knew and always reminded themselves that their behavior was not against public morality.

The survey on the attitude towards false information shows that when someone maliciously spreads bad information on the Internet, 42% of farmers ignore the information; 56% of farmers would warn their family and friends not to believe similar information; Not only will 2% of farmers not believe this information, but also report it to the police or relevant departments to find out and stop its continuous dissemination.

4 Conclusions and Suggestions

4.1 Conclusions

Through the investigation of farmers' information security literacy in five provinces, it is found that farmers' information security literacy is not high as a whole. Young people aged 25–35 who have received education above junior high school have a higher awareness of information security, while other groups have a weaker awareness of information security. Lack of information security knowledge is common. There are obvious differences between men and women in information security skills in the use of new media, and information security ethics need to be further improved. With the popularity of new media, it is urgent to improve farmers' information security literacy.

(1) For information security awareness

We investigated and analyzed the risk of personal information disclosure when respondents use new media, the importance of security software, and the attention to information security policies and regulations. The results show that there are significant differences between gender and information security awareness of new media users. The age and education level of the respondents will also affect the information security awareness. Young people aged 25–35 and farmers with junior middle school education or above have relatively good information security awareness. On the whole, farmers' awareness of information security needs to be further improved. Having a good awareness of information security is the premise of protecting personal information.

(2) Information security knowledge

We have conducted relevant analysis on mobile payment, anti-virus software, social media chat and information security laws and regulations. The results show that the overall situation of this part is general. Taking the mastery of network security and its legal knowledge as an example, 62% of farmers do not know the knowledge of network security and its law at all; 35% of farmers know this; 4% of farmers know very well. It shows that it is necessary to strengthen the cultivation of farmers' network security and legal knowledge. The results show that the popularization of information security knowledge in rural areas with relatively backward economic, cultural and technological development is not comprehensive. At the same time, the respondents have a low grasp and attention to "laws and policies related to personal information security".

(3) Information security capability

The overall situation of this part is inferior to that of consciousness. We have made relevant analysis on the problems of respondents clicking on unknown links, installing security software on mobile phones, data backup, closing location, granting the permissions required by applications and so on. The results show that 60% of the respondents will open unknown links and download the applications recommended by the system;

80% of the respondents' mobile location is turned on and will agree to the permissions required by the system. Gender differences have a greater impact on the ability of information security, and men have a higher ability of information security than women.

(4) Information security ethics

With its unique communication mode and inherent particularity, new media has changed the way of human life. On the other hand, it also produces a lot of unhealthy information. It is obvious that unhealthy information has a negative impact on cultural progress and then mislead values. The survey results of ethics in the use of new media show that 52% of the respondents do not know that all their speech acts on the Internet are monitored; 41% of the respondents knew it, but it was not serious; Only 22% of the respondents knew and strictly regulated their words and deeds. Generally speaking, farmers' information security ethics still need to be strengthened and improved.

4.2 Suggestions

In view of the current situation that farmers' information security literacy is not high, create an environment for improving information security literacy from the perspectives of farmers, grass-roots organizations, government and media, and innovate the channels and contents of information security knowledge training with short video app as the carrier to ensure the overall improvement of farmers' information security literacy.

(1) Enhance farmers' attention to information security literacy

On the one hand, farmers have not been exposed to the policy publicity of information security. The long-term peace has weakened people's awareness of the risk of information security, and they do not realize that the reason for the network fraud around them is the problem of information security. On the other hand, many people will think that information security is the business of the national information security department or confidentiality personnel and technicians, which has nothing to do with themselves, and they lack the understanding of the importance of information security to themselves. Therefore, farmers need to improve their information security literacy, enhance their awareness of information security, obtain knowledge about information security through multiple channels, improve the use skills and information processing ability of digital media, and standardize their network behavior with legal and moral standards.

(2) Improve the content and form of information security literacy training

The content of traditional information security literacy training mostly stays in the content of confidentiality and security and the theory of information security technology, such as military security, sovereign security, anti espionage, hacker attack and defense, cryptography theory, etc. the specific examples are mostly intelligence, espionage, network hacker and so on. There are fewer information security cases that are close to people's daily life, and less content accepted by farmers. This may be more serious. They believe that whether they have information security awareness is not important.

Information security is not related to their own responsibilities and obligations. Information security has nothing to do with current economic security, cultural security, scientific and technological security and other non-traditional security. At the same time, the training form is single and boring, which is limited to the preaching and publicity of relevant documents and regulations, which will only make people think that the universal education of information security awareness and literacy is a simple classroom knowledge teaching and publicity, which is just as important as fire prevention and anti-theft, resulting in disgust and exclusion and reducing the expected training effect.

Short video is a new way of Internet content dissemination. It is generally a video with a duration of less than 5 min spread on new Internet media. Compared with the traditional course teaching form, its duration is shorter and the content is more concentrated, which is convenient for the audience to focus on receiving enough valuable information in the fragmented time. Mobile short video attracts young people with its fine and compact features, attracting them from communicators to producers. As of December 2021, the utilization rate of medium and short video users among Internet users was 90.5% and the number of users reached 934 million. As a new thing, new media has gone deep into the countryside. Providing information security learning resources in the form of farmers' interest will greatly improve the effect of information security literacy education.

(3) Establish a multi-level national information security literacy training system

As a special education content, information security education should be vigorously advocated and supported from the government level, giving full play to the propaganda role of the mainstream media, especially the role of WeChat official account and micro-blog, so as to carry out nationwide publicity and education, so that the concept of information security is firmly rooted in people's brain. Bring individuals, grass-roots organizations, media institutions and state organs into the scope of information security literacy education, and develop targeted and cohesive training plans.

For example, some villages have not yet recognized the importance of information security literacy training, or think that information security literacy training is a matter of the state, has nothing to do with the village committee, lacks education and guidance on information security literacy for villagers, and does not make villagers aware of the importance of information security. In fact, both individuals, organizations, media and countries are the main body of information security. The improvement of farmers' information security literacy is inseparable from the power of participants. Therefore, from a macro perspective, to improve farmers' information security awareness, increase information security knowledge, improve information security ability, and abide by the ethical and legal bottom line in the use of new media, each participant needs to play a role together.

5 Conclusion

Information plays a more and more important role in Rural Revitalization and rural digital governance. Farmers' information security literacy determines the degree and process of rural information development. In view of the current situation that farmers' information

security literacy is generally not high, it is proposed to work together from farmers, government subjects, grass-roots organizations and media to improve information security awareness, disseminate information security knowledge, improve information security ability, and abide by ethics in information communication. The research enriches the research objects and categories of information security in theory, and provides ideas for the all-round development of farmers and rural areas.

References

1. Behrens, S.J.: A conceptual analysis and historical overview of information literacy. *Coll. Res. Libr.* **55**(4), 87–97 (1994)
2. American Library Association. Presidential Committee on Information Literacy. Final Report [EB/OL]. http://www.ala.org/ala/mgrps/divs/acrl/publications/whitepapers/president_ial.cfm. Accessed 8 July 2011
3. Allam, S., Flowerday, S.V., Flowerday, E.: Smartphone information security awareness: a victim of operational pressures. *Comput. Secur.* **42**(5), 56–65 (2014)
4. Samonas, S., Dhillon , G., Almusharraf, A.: Stakeholder perceptions of information security policy: analyzing personal constructs. *Int. J. Inf. Manag.* **50**(2), 44–154 (2020)
5. Schuster, E.: Building Ablockchain-based decentralized digital asset management system for commercial aircraft leasing. *Comput. Ind.* **126**(3),103393 (2021)
6. Ngoqo, B., Flowerday, S.V.: Information security behaviour profiling framework (ISBPF) for student mobile phone users. *Comput. Secur.* **53**(9), 132–142 (2015)
7. Wu, Y.Q., Zhang, X., Sun, H.B.: A multi-time-scaleautonomous energy trading framework within distributionnetworks based on blockchain. *Appl. Energy* **287**(1), 116560 (2021)
8. Cram, W.A., et al.: Organizational information security policies: a review and research framework. *Eur. J. Inf. Syst.* **26**(7), 605–641 (2017)



A Torque-Current Prediction Model Based on GRU for Circumferential Rotation Force Feedback Device

Zekang Qiu¹, Jianhui Zhao¹(✉), Chudong Shan¹, Wenyuan Zhao², Tingbao Zhang², and Zhiyong Yuan¹

¹ School of Computer Science, Wuhan University, Wuhan, China

qiuzekang@whu.edu.cn

² Zhongnan Hospital of Wuhan University, Wuhan, China

Abstract. Interventional surgery has many advantages over traditional surgery. However, plenty of preoperative training and animal experiments are needed to accumulate operation proficiency due to its high complexity. Virtual interventional surgery system provides a convenient and practicable way to gain experience for medical staff. In virtual interventional surgery, it is very important to generate accurate feedback force in order to obtain a better sense of realism and immersion. Our group have designed an electromagnetic force feedback device for virtual interventional surgery, which is based on the magnetic levitation principle to create circumferential rotation haptic force. The magnetic field is generated by the current passing through the coils. Therefore, it is very important to accurately control the value of the current. In our group's force feedback device, the performance of current generating method is not good enough. We observe that there is a temporal relationship in the data. Inspired by this, we proposed a Torque-Current Prediction Model based on GRU to predict the current value. The main process of the method is as follows: Firstly, find two known context data for the target data, so that sequence data can be formed. Next, model the relationship between the data in the sequence. Finally, the predicted current value is calculated based on this relationship and the known contextual data. We have conducted comprehensive experiments, and experimental results show that the proposed method outperforms previous method.

Keywords: Virtual interventional surgery · Electromagnetic force feedback · Torque-current prediction · GRU

1 Introduction

The surgery that introduces special guide wire and catheter into the human body to diagnose and locally treat the body disease according to the digital subtraction angiography is interventional surgery [12]. With its wide application, minimal invasion, obvious effect and minor complication, interventional surgery can fix shortcomings of traditional surgery [9]. Thus, it's a burgeoning method to treat cardiovascular and tumor diseases.

However, plenty of preoperative training and animal experiments are needed to accumulate operation proficiency due to its high complexity. As a result, virtual interventional surgery system provides a convenient and practicable way to gain experience for medical staff, especially the systems which depends on the force haptic feedback generation technology. However, it is crucial for the virtual system to generate the force feedback which includes the axial push-pull force of the catheter and the circumferential rotation force during the twisting process of the guide wire real-timely and accurately.

In 2000, CIMIT research center of United States developed a cardiovascular interventional surgery training equipment ICTs [3], and proposed essential components in virtual interventional surgery, but it lacks specific method to generate the key force feedback. In 2012, Li et al. [7] designed a catheter surgery training system, and conducted relevant research on blood vessel and surgical instrument modeling. In 2016, Guo et al. [5] studied the virtual force feedback in robot intervention surgery system. In 2018, Omisore et al. [10] simulated the wire catheter in interventional surgery. In 2020, Chen et al. [1] designed a clamping mechanism for circumferential force feedback device.

However, in the above research, the force feedback module is mostly designed by contacting mechanical pulley. Compared with mechanical contact force feedback, the electromagnetic non-contact method can make precise control of force feedback by completely avoiding mechanical friction [6]. Furthermore, it can vividly reproduce clinical operation mode of interventional surgery due to its unrestricted moving in the operation range [8]. The key principle of the method is to utilize controllable current to generate a specific magnetic field, with which the instrument equipped with magnetic material engenders force feedback.

Our group have designed an electromagnetic force feedback model for virtual interventional surgery [13]. To precisely control the value of the generated force requires precise control of the value of the current. In the device previously designed by our research group, the current value is predicted by a simple neural network model for a given angle and torque value. The neural network model is composed of a fully connected neural network with two hidden layers and a GRNN network. Although the accuracy of this method achieves an acceptable result, there is still room for improvement. We observe that there is a temporal relationship in the data. Inspired by this, a GRU-based time series modeling method is proposed to predict the current value.

The main contributions of this paper are as follows:

- We propose a data processing method that processes data into time series form.
- We propose a torque-current prediction model based on GRU for circumferential rotation force feedback device.
- We conduct comprehensive experiments, and the experimental results show that the proposed method has high performance.

2 Related Work

2.1 Force Feedback Interaction Model

The architecture of our research group's force feedback interaction model [13] is presented in Fig. 1. The interaction model's topology of electromagnetic coil array as shown

in Fig. 1a. Four identical electromagnetic coils are placed on the same plane. The structure has high symmetry and flexibility, can adjust the current in any combination of electromagnetic coils according to the force feedback requirements. The corresponding surgical instrument model is shown in Fig. 1b. The long cylindrical permanent magnet that two bottom surfaces are magnetized can provide the circumferential degree-of-freedom of the surgical instrument, and it interacts with the specific magnetic field excited by the coil array in Fig. 1a to generate the circumferential rotation force feedback. When the body center of long cylindrical permanent magnet coincides with the origin of *XYZ* coordinate system as in Fig. 1c, the force feedback interaction model can generate force feedback with its working mechanism. In addition, θ is the angle between the N-pole of the cylindrical permanent magnet and the positive direction of *X* axis.

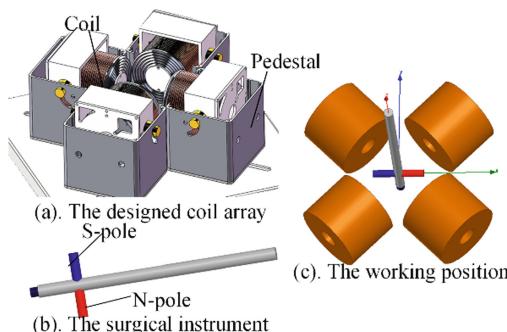


Fig. 1. Details of force feedback interaction model and their working position.

According to the motion principle of rigid body rotating around one fixed axis, two requirements need to be met for operator to obtain the circumferential rotation force feedback:

- The electromagnetic resultant force on the long cylindrical permanent magnet is 0, which ensures that the permanent magnet will not produce any movement except for the circumferential rotation around the shaft, so as to ensure the purity of the circumferential rotation force feedback.
- The rotation torque of the long cylindrical permanent magnet around the fixed axis is not zero and the value is considerable, which can ensure the effectiveness of the circumferential rotation force.

In order to meet the above requirements, combined with the magnetic field distribution characteristics of the electrified coils and the highly symmetrical topology of 4-coil array, the basic operation principle for above-mentioned model is as follows: the coil pairs with diagonal positions must be given the same excitation current in the same direction.

Based on the basic operating principle of the model, in order to generate real-time and accurate circumferential rotational force feedback, an optimal current distribution

strategy is applied in the force feedback model. The characteristics of the strategy are as follows:

- Efficiency: Choosing the pair of coils, which results larger torque when powering the same value of exciting current.
- Simplicity: The whole working area of the model can be mapped to the minimum symmetric interval by using the characteristics of relative position and absolute position of the topological structure of the coil array.

According to the symmetry of coil array, the whole rotation range of the permanent magnet can be divided into eight arc regions with 45° angle. The numerical characteristics of the magnetic field excited by the current in each region are exactly the same. Thus mapping between the corresponding regions can be carried out to reduce the amount of data calculation. Based on the above two characteristics of the optimal current allocation strategy, the optimal current distribution strategy can be expressed as the following formulas: for any angle θ of the permanent magnet, a 3D tuple (x, y, z) is proposed to determine optimal current allocation in the electromagnetic coil array:

$$x = \begin{cases} \left[\frac{\theta}{90} \right] \text{mod} 2 & 3 < y < 45 \\ \neg\left(\left[\frac{\theta}{90} \right] \text{mod} 2 \right) & y < 3 \end{cases} \quad (1)$$

$$y = |45 \times \left(\left[\frac{\theta}{45} \right] \text{mod} 2 \right) - \theta \text{mod} 45| \quad (2)$$

$$z = \left[\frac{\theta}{180} \right] \quad 0 < \theta < 360 \quad (3)$$

In the 3D tuple, x decides which electromagnetic coil pair to apply current, and mod is the calculation of remainder. When its value is 0, it means that 1, 3 coils work; when the value is 1, the 2, 4 coils work. While y is the angle of equivalent mapping in the minimum symmetric interval, and the value is 0° - 45° . We use z to indicate the current direction of the coil pair, and the value is 0 or 1. Each corresponding unique triplet represents a certain optimal current allocation strategy.

2.2 RNN and GRU

RNN (Recurrent Neural Network) is a type of recurrent neural network that takes sequence data as input, performs recursion in the evolution direction of the sequence, and connects all nodes (recurrent units) in a chain [4]. RNN can model the relationship between sequence data, so RNN are often used to deal with tasks in which the data is in the form of sequences, such as natural language processing, machine translation, and speech recognition.

The most basic RNN unit remembers all the information in the sequence, so when the sequence length is long, the problem of gradient disappearance and gradient explosion may occur. LSTM (Long Short-Term Memory) [11] is a variant of RNN unit. Its basic idea is to introduce a “gating device”, which can effectively solve the problems of gradient disappearance and gradient explosion.

Although LSTM can solve the problem of gradient disappearance and gradient explosion caused by long-term dependence of recurrent neural network, LSTM has three different gates and many parameters, making it more difficult to train. Later, another recurrent neural network unit, GRU (Gate Recurrent Unit) [2], appeared. GRU only contains two gated structures, and when all hyperparameters are tuned, GRU and LSTM perform equally well. The GRU structure is simpler, easier to train, and faster in operation.

3 Data and Method

This paper proposes a torque-current prediction model based for circumferential rotation force feedback device. The torque-current prediction model is used to calculate the current of the coil array. It has two inputs, one is the angle between N-pole of surgical instrument model and the x-axis, and the other is the target moment. And its output is the value of current.

3.1 Potential Sequence Characteristics in Data

Data in the solver of Ansoft Maxwell software are collected, and irrelevant items are eliminated by principal component analysis to form an effective and reasonable dataset. The model parameter settings of Ansoft software are shown in Table 1. Each data is composed of coil cross-section current which is calculated by multiplying single wire current and number of windings, torque of permanent magnet rotating in the positive direction of Z axis and the angle between N-pole of permanent magnet and positive direction of X axis. The training dataset consists of 966 items with θ in range of 0° – 45° and current in range of 0–3680 A. After obtaining these data, the angles, torques, and currents in the data are normalized separately.

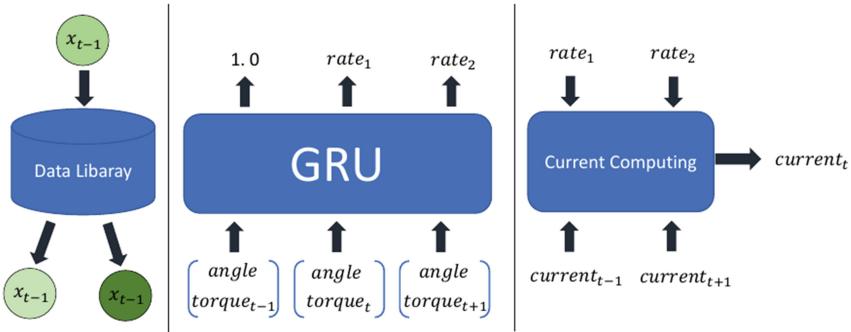
We found that the data in the dataset have the characteristic that the current value increases with the increase of the torque value when the angles are the same. Therefore, we consider that there is a serial relationship between the data, and we will make full use of this relationship in the torque-current prediction model to improve the accuracy of the predicted current.

3.2 Torque-Current Prediction Model Based on GRU

Each piece of data in the training set described in Sect. 3.1 consists of an angle value, a torque value, and a current value. In this section, we denote each piece of data as a data unit. The input to Torque-Current Prediction Model is an angle value $angle$ and a torque value $torque_t$, and the output is the current value $current_t$ predicted by the model. The whole process of the current prediction model is shown in Fig. 2, and the whole process can be divided into three stages.

Table 1. The model parameter settings of Ansoft software

Parameters	Value	Parameters	Value
Inner radius of coil(mm)	13.5	Radius of magnet(mm)	6
Outer radius of coil(mm)	35	Height of magnet(mm)	36
Height of coil(mm)	62	Material of magnet	NdFe35
Num of windings of coil	1840	Material of coil	Copper
Distance of coil set(mm)	84	Wire gauge of coil(mm)	0.81

**Fig. 2.** The overall processing flow of the proposed Torque-Current Prediction Model Based on GRU.

In the first stage, two data units are found in the known dataset based on the input $angle$ and $torque_t$. The first one is the data unit $x_{t-1} = (angle, torque_{t-1}, current_{t-1})$ whose angle value is the same as $angle$ and the torque value is slightly smaller than $torque_t$, while the second one is the data unit $x_{t+1} = (angle, torque_{t+1}, current_{t+1})$ whose angle value is the same as $angle$ and the torque value is slightly larger than $torque_t$. It should be noted that the known dataset here is the training set described in Sect. 3.1, and the input $angle$ and $torque_t$ are the data in the testing set.

In the second stage, we use a GRU network to model the relationship between data units. In this paper, the sequence length of the GRU network is 3. Each element in the input sequence is a feature vector of dimension 2 formed by the angle value and the torque value, and the three elements of the input sequence are respectively composed of the angle value and the torque value in x_{t-1} , x_t , x_{t+1} . Each element in the output sequence of the GRU is a feature vector with a dimension of 6, and each feature vector is mapped to a value through a linear layer, which represents the ratio of the current value of the data unit to the current value of the previous data unit. We denote the values of the second and third feature vectors in the output sequence mapped by the linear layer as $rate_1$ and $rate_2$, respectively. $rate_1$ and $rate_2$ represent the results of $current_t/current_{t-1}$ and $current_{t+1}/current_t$ predicted by GRU, respectively.

In the last stage, $current_t$ is calculated according to $rate_1$ and $rate_2$ output by GRU, as well as $current_{t-1}$ and $current_{t+1}$. The calculation process is as follows.

$$current_t = \frac{1}{2}(current_{t-1} \times rate_1 + current_{t+1} \times rate_2) \quad (4)$$

It should be noted here that $current_{t-1}$ and $current_{t+1}$ are known data, so they can be used directly. The $current_t$ is the predicted current value corresponding to the input angle and torque, that is, the output value of the Torque-Current Prediction Model.

3.3 Training the GRU Model

In order for the Torque-Current Prediction Model to work effectively, the GRU module of the Torque-Current Prediction Model needs to be trained. As explained in the previous section, the input sequence length of GRU in Torque-Current Prediction Model is 3, and each element in the sequence is a vector of dimension 2 composed of angle and torque. And these three vectors need to meet the conditions: the angle is the same and the torque is sequentially increasing. The 966 pieces of data in the training data set mentioned in Sect. 3.1 are independent of each other. In order to train the GRU, the independent data in the training set need to be processed into data that satisfy the GRU input sequence form.

The specific processing flow is as follows: First, traverse 966 pieces of data, and put the data with the same angle value into the same list. Since the angle value ranges from 0 to 45, 46 lists are obtained finally. Next, sort the data in each list in order of increasing torque value. Finally, a list L is randomly selected, and a piece of data x_t in L is randomly selected. After obtaining x_t , take the data x_{t-1} that is in front of x_t and spaced by m data from x_t in L , and similarly take the data x_{t+1} that is behind x_t and spaced from x_t by n data in L . The angle and torque values of x_{t-1} , x_t , and x_{t+1} respectively form three vectors of dimension 2. The sequence of these three vectors is a legal input sequence of the GRU in the Torque-Current Prediction Model. The output sequence (ground truth) corresponding to the input sequence is $(1.0, rate_1, rate_2)$, where $rate_1$ is the ratio of the current value of x_t to the current value of x_{t-1} and $rate_2$ is the ratio of the current value of x_{t+1} to the current value of x_t . Each piece of data and the corresponding ground truth in the training phase are obtained by this process. In addition, the random intervals m and n in the above process range from 0 to 3. Random intervals can increase the diversity of the data, thereby enhancing the generalization ability of the model.

During training, the optimizer used is Adam, the loss function is MSE (mean squared error), and the learning rate is 0.01.

4 Experiments

4.1 Obtaining Testing Dataset

In order to verify the performance of Torque-Current Prediction Model, we collected 15 pieces of data that are within the reachable range and do not coincide with the training set through the simulation of Ansoft Maxwell software as the test set. Similarly, these

15 pieces of testing data are also independent data, which do not conform to the input format of GRU in Torque-Current Prediction Model, so it needs to be further processed into the form of input sequence that satisfies GRU.

The specific processing flow is as follows: First, for each of the 15 pieces of data x_t , in the 46 lists obtained in Sect. 3.3, find a list L whose angle value is the same as the angle value of x_t . Next, traverse from back to front in L , take the first data with a torque value less than x_t as x_{t-1} , and traverse from front to back and take the first data with a torque value greater than x_t as x_{t+1} . Finally, use the angle and torque of x_{t-1} , x_t , x_{t+1} to form three vectors of dimension 2, which is the input of the GRU in the testing phase. This process is mainly to find two contextual data for x_t in the known data, so that x_t can be processed into data in the form of a sequence.

4.2 Evaluation Metrics

The Torque-Current Prediction Model proposed in this paper is a regression task, so this paper will adopt the metrics commonly used in regression models to evaluate the proposed method to verify its performance, including RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error). The three indicators are defined as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (6)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (7)$$

Where n is the number of test data, where $n = 15$; \hat{y}_i is the predicted current value, and y_i is the corresponding actual current value.

4.3 Experimental Results

The experimental results are shown in Fig. 3. It can be seen that there is basically no deviation between the current value predicted by the Torque-Current Prediction Model and the real current value.

The current prediction model based on GRU proposed in this paper is more accurate than the original model based on the fusion of BPNN and GRNN. The comparison of evaluation metrics results is shown in Table 2. It can be seen that the GRU-based Torque-Current Prediction Model proposed in this paper is better than the original method in all evaluation metrics. In addition, we also performed some ablation experiments. In order to find the best RNN unit, we tried to replace the GRU unit with a normal RNN unit and LSTM unit, respectively. The evaluation metrics results of different RNN units are also shown in Table 2.

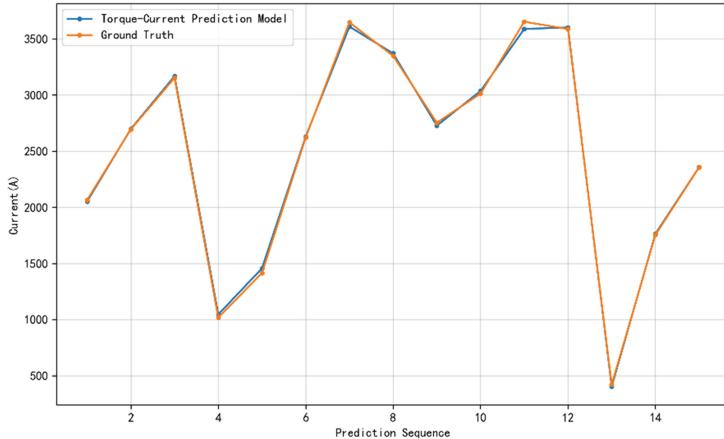


Fig. 3. Current Prediction Results of Torque-Current Prediction Model on Testing Set.

Table 2. Evaluation Metrics of Different Methods

Methods	RMSE	MAE	MAPE
MLP + GRNN	45.54	36.07	1.71%
RNN	42.75	34.91	1.72%
LSTM	29.29	22.79	1.28%
GRU	26.81	21.55	1.16%

5 Conclusion

In this paper, we propose a Torque-Current Prediction Model based on GRU. This model is used for current value prediction for our group's Circumferential Rotation Force Feedback Device. The proposed method achieves high-performance current value prediction by serializing the data and modeling the relationship between the sequence data. We have conducted comprehensive experiments, and the experimental results show that the method proposed in this paper exceeds the original current value prediction method of the force feedback interaction model. The focus of this paper is to study the method of high-performance current value prediction, and the real-time performance has not been studied for the time being. In future work, we will study the current prediction methods with both high accuracy and high real-time performance.

Acknowledgements. This work was supported by the Natural Science Foundation of China under Grant No. 62073248, the Translational Medicine and Interdisciplinary Research Joint Fund of Zhongnan Hospital of Wuhan University under Grant No. ZNJC201926.

References

1. Chen, Z., Guo, S., Zhou, W.: A novel clamping mechanism for circumferential force feedback device of the vascular interventional surgical robot. In: 2020 IEEE International Conference on Mechatronics and Automation (ICMA), pp. 1625–1630. IEEE (2020)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078) (2014)
3. Dawson, S.L., Cotin, S., Meglan, D., Shaffer, D.W., Ferrell, M.A.: Designing a computer-based simulator for interventional cardiology training. *Catheter. Cardiovasc. Interv.* **51**(4), 522–527 (2000)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
5. Guo, J., Yu, Y., Guo, S., Du, W.: Design and performance evaluation of a novel master manipulator for the robot-assist catheter system. In: 2016 IEEE International Conference on Mechatronics and Automation, pp. 937–942. IEEE (2016)
6. Kim, Y., Parada, G.A., Liu, S., Zhao, X.: Ferromagnetic soft continuum robots. *Sci. Robot.* **4**(33), eaax7329 (2019)
7. Li, S., et al.: A catheterization-training simulator based on a fast multigrid solver. *IEEE Comput. Graphics Appl.* **32**(6), 56–70 (2012)
8. Li, X., Yuan, Z., Zhao, J., Du, B., Liao, X., Humar, I.: Edge-learning-enabled realistic touch and stable communication for remote haptic display. *IEEE Network* **35**(1), 141–147 (2021)
9. Murali, N., Ludwig, J.M., Nezami, N., Kim, H.S.: Oligometastatic disease and interventional oncology: rationale and research directions. *Can. J.* **26**(2), 166–173 (2020)
10. Omisore, O.M., et al.: Towards characterization and adaptive compensation of backlash in a novel robotic catheter system for cardiovascular interventions. *IEEE Trans. Biomed. Circuits Syst.* **12**(4), 824–838 (2018)
11. 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. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
12. Siracuse, J.J., Farber, A.: Lower extremity vascular access creation is a marker for advanced end-stage renal disease. *J. Vasc. Surg.* **71**(6), 2185 (2020)
13. Zhao, J., Lin, Y., Yuan, Z.: Designing and simulation of electromagnetic force feedback model focusing on virtual interventional surgery. *Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/J. Comput.-Aided Des. Comput. Graphics* **33**(8), 1254–1263 (2021)



Development and Application of Augmented Reality System for Part Assembly Based on Assembly Semantics

Yingxin Wang¹, Jianfeng Lu^{1,4(✉)}, Zeyuan Lin², Lai Dai¹, Junxiong Chen³, and Luyao Xia^{1,4}

¹ College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

lujianfeng@tongji.edu.cn

² School of Software Engineering, Tongji University, Shanghai 201804, China

³ School of Automotive Studies, Tongji University, Shanghai 201804, China

⁴ Engineering Research Center of Enterprise Digital Technology, Ministry of Education, Shanghai 201804, China

Abstract. Currently, most assembly guidance augmented reality systems focus on assembly geometry relationship using traditional interaction methods, leading to poor interactivity. To improve the situation, firstly, an augmented reality system framework for part assembly based on assembly semantics is proposed. Secondly, the intelligent extraction of semantic information of assembly is realized, based on SolidWorks secondary development technology. Then the granularity modeling problem of the assembly scenario is studied. And finally, an augmented reality system for part assembly based on assembly semantics is developed, which verifies the effectiveness of the assembly guidance based on assembly semantics, and improves the human-computer interaction.

Keywords: Augmented reality · Assembly semantics · Assembly guidance · Human-computer interaction

1 Introduction

Traditional workshops have long assembly processes and standardized assembly requirements, meanwhile, assembly operations are still inseparable from manual operations. However, relying on manual operation, the problems in the traditional assembly workshop, such as extended time, low assembly efficiency, and high error rate [1] remain unresolved, especially when the requirements of product design are gradually increasing. Hence, the effect of augmented reality (AR) technology in assembly operations has received extensive attention. Augmented reality can combine virtual information with the physical world to expand and enhance the physical world [2]. Therefore, AR-based assembly guidance can present a virtual-real fusion scene to the operator through related equipment.

There has been considerable research on the application of AR for assembly guidance. However, these AR assembly guidance technologies primarily focus on the geometric relationship of assembly and rarely involve the semantics relationship of assembly. Semantic information can describe significant information such as part information, which plays an important role in assembly guidance. When AR is used for assembly operations, assembly information obtained from the CAD system needs to be centered on human operations, converting the part information and assembly process into assembly semantics, thus making it more convenient for assembly personnel to operate [3].

In addition, there are various forms of interaction in AR assembly guidance, such as haptic interaction [4] and voice interaction [5]. Among these, gesture-based interaction is recognized as one of the criteria for AR-assisted assembly [6]. But gesture recognition is also implemented in various ways, such as glove-based, sensor-based, radar-based [7], etc. Most of them require multi-device support. We use the latest gesture recognition technology, which requires only the phone camera to complete the interaction.

The paper realizes assembly guidance function based on AR technology and assembly semantics. It expounds on the research background and the framework of the system. Then it clarifies the generation process of assembly semantics information. The paper also introduces the research on the assembly situation granularity model and the specific implementation information of the system.

2 The Overall Framework of the Augmented Reality System for Parts Assembly

The AR system for parts assembly based on assembly semantics first reads data from a model and obtains the semantic model of assembly; then, it establishes the granularity model of the assembly context based on the semantic model and divides the assembly steps to match and identify the assembly context in the actual process; then, it visualizes the assembly induced information based AR; finally, it provides real-time assembly guidance for the assembly process in the mobile application through gesture-interaction.

The general framework of the AR system for parts assembly consists of an assembly semantics generation module, an assembly context granularity modelling module, a guiding information visualization module, and a mobile device module (see Fig. 1).

2.1 Assembly Semantics Information Generation Module

Assembly semantics information integrates the physical information of each part and the assembly information between parts. By classifying the geometric rules of the part and the assembly relationship between the parts, assembly information of the part can be described uniformly. The assembly semantics generation module uses the secondary development technology of SolidWorks to abstract and summarize the semantics information of the part from the 3D model file of the pre-processed part. This module gives full play to the benefits of the CAD software and provides the basis for establishing assembly context.

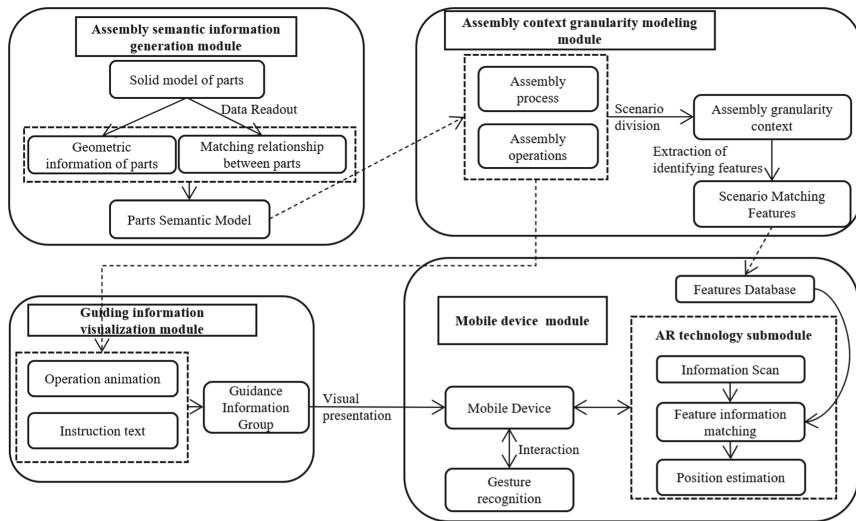


Fig. 1. The overall framework of the augmented reality system for part assembly based on assembly semantics

2.2 Assembly Context Granularity Modeling Module

The assembly context granularity modeling module serves to establish and divide the assembly context granularity model based on the assembly information to provide identification data and guidance information for assembly guidance. The module first determines the constraint relationships and assembly methods between parts and components, obtains the semantics model of parts, and performs modular division of possible assembly combinations; then it establishes the hierarchical model of assembly components and completes the granularity division of assembly scenarios; it further simplifies the assembly steps according to the granularity transformation of assembly scenarios, and eventually determines the divided assembly context groups.

2.3 Guiding Information Visualization Module

The guiding information visualization module gives visual information that can be presented in an AR scene. AR technology loads virtual content into the actual scene for combined display and therefore offers the information more intuitively. To perform this function of AR, it is necessary to transform the information that guides the assembly into visual information which can intuitively guide the assembly. Therefore, based on the scenarios obtained in the assembly context granularity modeling module and the pointing relationships between the scenarios, the system generates visual guidance information including instructional text and assembly animations.

2.4 The Mobile Device Module

The mobile device module is a module that displays assembly guidance information and interacts with the assembler during the assembly process. This module is based

on a mobile device that contains a camera that scans the environment and a database containing assembly scenario information. When the assembly guidance is at the open state, the matching recognition module uses the camera to scan the assembly environment and get the data to match the assembly context database to find the matching scenarios. After identifying the scenario, the visualization information from the induced information visualization module is called for providing the assembly guidance. At the same time, the module interacts with the assembler through gestures to enable functions such as turning on or off the assembly guidance of the part and displaying the overall part information, which is a more natural interaction process than the traditional touch method.

3 Assembly Semantics Information Generation

3.1 The Definition of Assembly Model Information

In this paper, the assembly model is considered as a combination of geometric model and semantics model (see Fig. 2). The geometric model includes the geometric model of the assembly and the geometric model of each part. Both are derived from the CAD model system and can be used as visual information in the AR system. The semantics model of the assembly includes the basic information of the assembly, the basic information of each part, and the assembly information between parts, which are combined to assist in guiding the assembly process.

Basic information of the assembly including the name of the assembly, overall dimensions, number of parts, and other information.

Basic information of each part including the name, ID, material, and configuration of each part.

Assembly information between parts includes the names of the parts and mate type, and the mating-type is classified concerning SolidWorks.

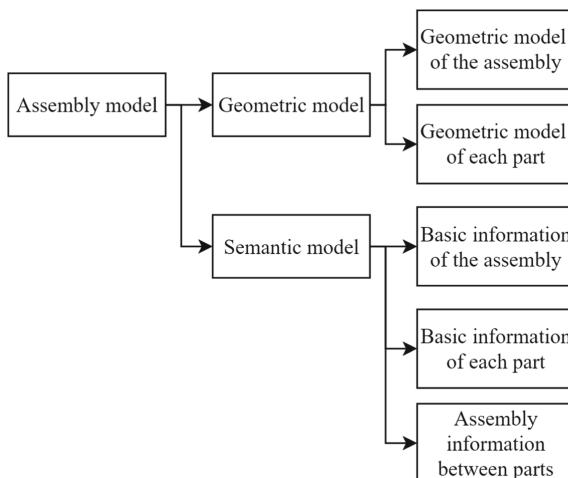


Fig. 2. The assembly model

3.2 Acquisition of Semantics Information in SolidWorks

CAD software can complete the solid modeling, and generate 3D model static files and animation files as visual information in the AR system. However, in the process of correct modeling, the assembly information is already generated. For example, in SolidWorks, some information can be obtained without operations in the feature tree. You can directly view the name of the top-level assembly, names of components, the name of each component's material and other information from the feature tree. On the other hand, you can obtain some information only after some special operations. You can get the dimensions of a part or assembly after adding a bounding box to it.

However, as can be seen from the two types of information, assembly information in SolidWorks is rather fragmented, and the operations of SolidWorks are not accessible. So it is tedious to obtain assembly semantics information based on the assembly semantics model defined above. To extract the assembly semantics information accurately and automatically, the paper uses the secondary development technology of SolidWorks to generate the semantics information with a specific format, which provides the necessary basis for the subsequent development of the assembly-induced system.

3.3 Automatic Extraction of Assembly Semantics Information Based on SolidWorks

Dassault, the company that owns SolidWorks, offers recorded macros and APIs for secondary development. The API-based development of SolidWorks can be used to automate some SolidWorks design work [8] and read related objects and properties. This paper develops a desktop application independent of SolidWorks.

Automatic extraction of assembly semantics information is not constrained by the operations of SolidWorks. Firstly, connect with the open SolidWorks file and get the ongoing ModelDoc2 object. Secondly, determine whether the file type is an assembly document (AssemblyDoc). If so, travel through the relevant objects from top to bottom, based on the architecture of SolidWorks. Thirdly, call the API to get the relevant information. Last, generate the information with a specific format in the text file. This paper does not specify some other boundary conditions in the whole process.

In the program (Fig. 3), the “Connect” button is used to connect to SolidWorks, the “Load” button is used to extract and generate the assembly semantics information, and the “Help” button is used to show the usage of the software; The right box of the “Folder path” button can be used to input the path of the save file, the “Select” button is used to pop up the interface to select the path of the save file, and the “Save” button is used to save the file. The “Select” and “Save” buttons can only be used when there is assembly semantics information in the information box, otherwise, it will be greyed out.

The file level corresponds to the assembly semantics information level, consisting of Assembly Data, Components Data, and Mates Data. The hierarchy of file formats is shown in Fig. 4 below. A blank line separates the three levels, and the information under each type of information will have one more space than the first line of the corresponding level. The basic information of each part is distinguished by Name, and ID, Material Name, and Referenced Configuration will have one more space than the first line of

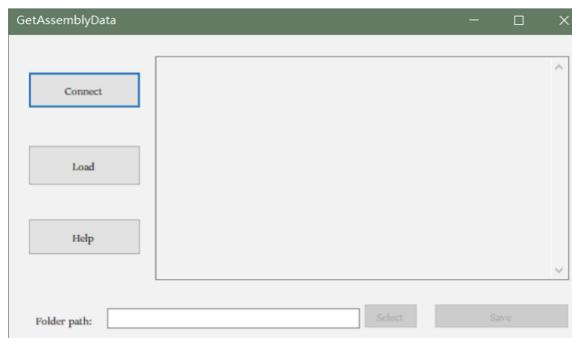


Fig. 3. The program interface

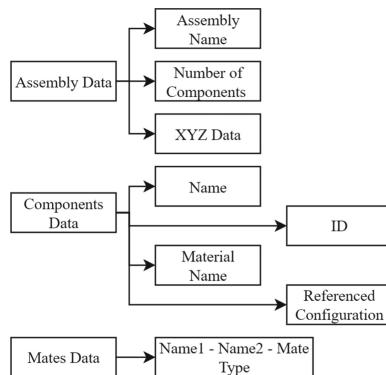


Fig. 4. The hierarchy of file format

Name. The format of the assembly information between parts is unified as follows: Name of Part1 - Name of Part2 - Mate Type.

With the .txt file, you can further refine the assembly induction information according to the assembly semantics, set the text box in the unity system, write automation scripts to set the text box text as information, select the .txt file in the unity interface, and use the text box content as virtual guidance information to join the assembly induction system and display it on the device.

4 Granular Modeling of the Assembly Context

Granular computing, through establishing effective computing models, can substitute satisfactory approximate solutions for accurate solutions, thus raising efficiency [9]. However, for multi-layer problems, the granular partition is required to establish multi-granularity models to meet the various need for accuracy of different layers, and finally optimise computing efficiency. In this chapter, the concept of the part group is established to simplify the structures of assembly models. Then the model of granular partition is restated, and its partition form is improved by providing a more precise approach

to partition sub-contexts. Finally, a novel granular model under the new sub-context partition is established.

4.1 Modular Partition of Assembly Steps

For complicated mechanical structures, performing the assembly guidance for every single part may lead to a lengthy and complex guiding system and affect assembly efficiency due to users' inevitable frequent switches between reality and guidance. Therefore, before establishing the multi-layer granular model, the assembly steps should be modularized in advance, creating a context partition with a clear distinction between major and minor parts to raise the assembly efficiency.

It should be noted that in assembly practice, workers can usually infer the position and coordination of the fasteners by themselves. However, workers need additional instruction to figure out the functional and structural parts' position and coordination, which are mostly defined by the specific mechanical need. In this way, an approach to modularizing the assembly steps can be introduced.

First, a beginning part should be defined as a reference. This part should be functional or structural, usually the frame or body part. Beginning with the part, every newly added functional or structural part, together with all the fasteners that connect the new part with the assembled parts, is defined as a part group. In this way, an assembly step is defined as installing a new part group (see Fig. 5). This partition meets the workers' common cognition in assembly and can lower the difficulty of eviting chaos when instruction collides with users' expectations.

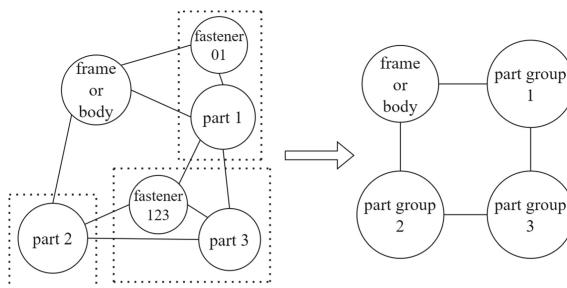


Fig. 5. Simplify assembly steps according to the part group

4.2 Granular Partition of Assembly Context

In a mechanical assembly context, the granular partition is defined as building the multi-layer model of parts. This paper refers to the granular partition by Xia from Tongji University [10], designs three granular layers from coarse to fine (see Fig. 6): The first layer is the context layer of part semantics, which consists of the information set of all parts' semantics. The second layer is the sub-context layer, the set, modularized by part groups, of part nodes and assembly semantics information. The third layer, the elemental layer, consists of the elemental attributes of the assembly, including the information of

parts like sizes, positions, and attitudes. This partition has the merits of corresponding with instinct, being transparent and brief, and fitting well with engineering software.

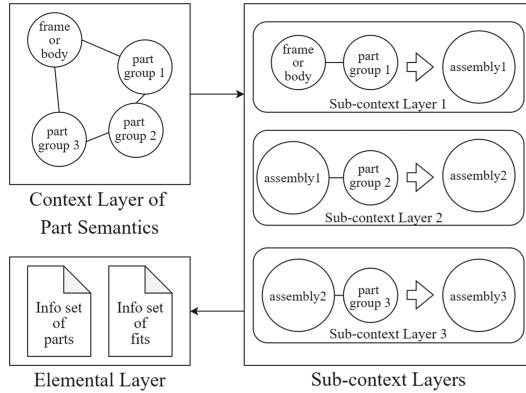


Fig. 6. Granular partition of assembly context

4.3 Granular Transformation in Task Contexts

Granular transformation is transforming from one granular level to another without altering the system function [11]. In the assembly context, granular transformation means the coarse granular transformation that modularizes the already assembled parts. In other words, the assembled parts will be viewed as a new ‘part’, ignoring the inner fit relationship. In this way, the assembly contexts and steps are simplified and the difficulty that computers recognize specific assembly contexts is eased. As shown in Fig. 7, the assembled part 2 can be simplified as a new ‘part’ 2^* with assembly semantics in a novel granular space.

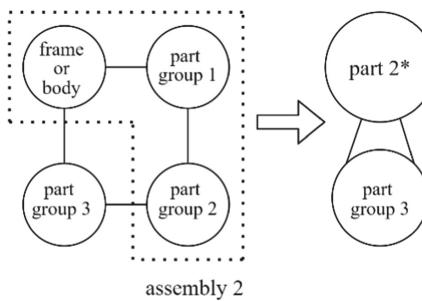


Fig. 7. Granular transformation in task contexts

5 Example System Implementation

In this paper, the simple press has been processed as the assembly to be guided, the semantics information of the assembly is extracted based on the content of Sect. 3;

the assembly process is simplified and determined based on the content of Sect. 4; the visualization content is determined based on the semantics information and its assembly process. The AR module is developed in the Unity system using Vuforia SDK, and the gesture recognition is implemented using the manomotion SDK.

5.1 Granular Transformation in Task Contexts

The simple press has seven parts, including one frame, one press sheet, one long bar, one short bar, and three sleeves. The specific assembly process is shown in Fig. 8 below.

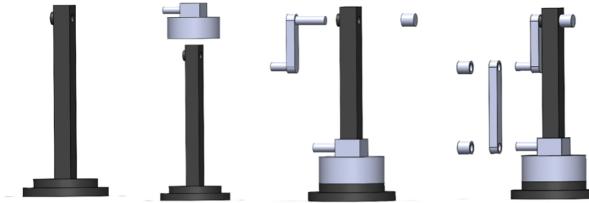


Fig. 8. The step 0–3 of the simple press

From this, the assembly relationship diagram of the press can be created and simplified by part group according to the method in Sect. 4.1 (see Fig. 9(1)). On this basis, it can be divided into sub-contexts according to the assembly of the part group. Finally, three sub-contexts are separated in order by combining the spatial and fit relationships of the parts (see Fig. 9(2)).

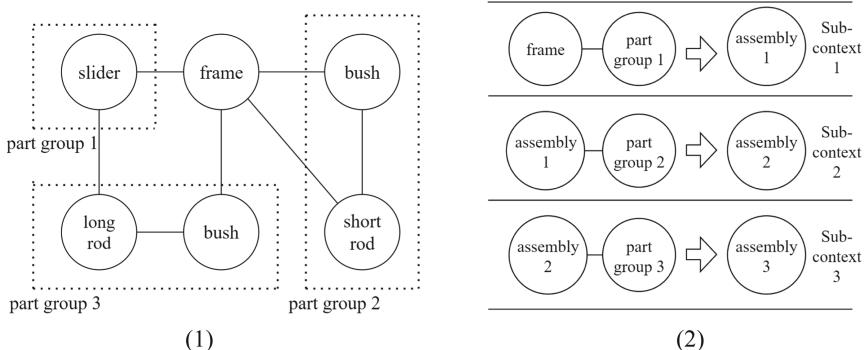


Fig. 9. Simplification and division of assembly scenarios

5.2 The Development of The AR Module

The development of the AR module starts with determining the visualization content to be displayed by the assembly guidance system. Considering the convenience and intuitiveness, the visualization content determined in this paper is static assembly semantics

information on the one hand and dynamic assembly step-by-step animation on the other. The assembly semantics information is extracted from the software in Sect. 3, and the operator can determine the assembly problem based on this content. In contrast, the dynamic assembly animation is based on the CAD model of the assembly, which is more intuitive than static pictures and models (see Fig. 10(1)), and is created by the TimeLine plug-in that comes with Unity.

Vuforia is an AR solution that supports AR program development for Android and iOS in the form of a Unity plugin. This article is based on implementing the program using C# language and deploying it to the Android platform for testing, with the following process:

Create a database on the Vuforia website, and scan the objects to be identified at each assembly step by the Object Scanner App provided by Vuforia. After that, upload the scanned data to the database. Once this is done, you can import the Vuforia development package into Unity and develop the objects you have created according to the visualization content you have determined.



Fig. 10. Information displayed by the device

5.3 The Development of the Gesture Recognition Module

In the actual assembly process, the conventional touch-based interaction is no longer applicable because the assembler needs his hands to perform the assembly; at the same time, the assembly is often carried out in a noisy environment and the voices of different assemblers may conflict, which makes the sound control may produce errors. So the system adopts the gesture control method to interact. For gesture recognition, the system is developed with manomotion's gesture recognition development interface.

The system mainly uses gesture interaction to control two functions: whether to turn on the AR assembly guidance and to view information about the parts of this assembly.

The function of whether to turn on augmented reality assembly guidance improves the autonomy of the assembler. In augmented reality guidance, as the assembly guidance information is presented directly to the view combined with the physical world, it may obstruct the view and cause trouble when the part assembly guidance is not needed. To address this issue, the system decides whether to turn on the guidance up to the assembler and takes the form of one quick tap of the index finger and thumb to turn augmented reality assembly guidance on or off.

The function to view the part information of this assembly allows the assembly to move away from the guidance of individual steps to view all the part information

and assembly information of the entire assembly process, thus better supporting the assembly process and clarifying the positioning of the current assembly step for this. The system takes a fist and then releases to view this assembly part information or return to single-step assembly guidance (see Fig. 10(2)).

6 Summary

Firstly, the paper proposes an augmented reality system for part assembly based on assembly semantics, consisting of the assembly semantics generation module, the assembly context granularity modeling module, the guiding information visualization module, and the mobile device interaction module, meanwhile illustrating the role of the four modules and their interconnections with each other. Secondly, in order to determine the assembly semantics accurately and conveniently, the paper implements the intelligent extraction of assembly semantics information based on SolidWorks' secondary development technology. Then, this paper studies the problem of granular modeling of assembly context. The assembly structure is simplified by the concept of the part group, and the granularity of assembly context is divided and transformed on this basis. Finally, on the Unity platform, the paper develops an augmented reality system for part assembly based on assembly context that uses gestures to interact, which confirms the feasibility of the proposed system framework.

Acknowledgments. Research work in this paper is supported by the National Natural Science Foundation of China (Grant No. 72171173) and Shanghai Science and Technology Innovation Action Plan (No. 19DZ1206800).

References

1. Li, W., Wang, J., Lan, S., et al.: Content authoring of augmented reality assembly process. *Comput. Integr. Manuf. Syst.* **25**(07), 1676–1684 (2019)
2. Zhao, X., Zuo, H.: Aviation applications and prospect of augmented reality. *Aviation Maintenance Eng.* **6**, 23–25 (2008)
3. Chen, C.J., Hong, J., Yang, G.Q.: Assembly and disassembly guiding system based on semantic context for complicated mechanical equipments. *Comput. Integr. Manuf. Syst.* **20**(7), 1599–1607 (2014)
4. Sun, M., He, W., Zhang, L., et al.: Smart haproxy: a novel vibrotactile feedback prototype combining passive and active haptic in AR interaction. In: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 42–46. IEEE (2019)
5. Dong, Q., Li, B., Dong, J., et al.: Realization of augmented reality assembly voice interaction for head-mounted glasses. *Manufact. Autom.* **42**(10), 77–80 (2020)
6. Agati, S.S., Bauer, R.D., Hounsell, M.S., et al.: Augmented reality for manual assembly in industry 4.0: gathering guidelines. In: 2020 22nd Symposium on Virtual and Augmented Reality (SVR), pp. 179–188. IEEE (2020)
7. Zhang, F.J., Dai, G.Z., Peng, X.: A survey on human-computer interaction in virtual reality. *Scientia Sinica Informationis* **46**(12), 1711–1736 (2016)
8. Zhang, J.W., Wang, B., Du, Y.J.: Three-dimensional parametric design on extruder frame structure. *Forging Stamping Technol.* **46**(01), 131–135 (2021)

9. Pang, J.F., Song, P., Liang, J.Y.: Review on multi-granulation computing models and methods for decision analysis. *Pattern Recognit. Artif. Intell.* **34**(12), 1120–1130 (2021)
10. Xia, L.Y.: Research and application of production visualization system based on augmented reality in digital workshop. Tongji University (2020)
11. Wu, Q., Wang, Y.F., Bian, J.N., et al.: Granularity transformations based on a new CDFG format for granularity selection in hardware-software partitioning. *J. Comput. Aided Des. Graph.* **17**(3), 387–393 (2005)



Research on Augmented Reality Assisted Material Delivery System in Digital Workshop

Zhaojia Li^{1(✉)}, Hao Zhang^{1,2}, Jianfeng Lu^{1,2}, and Luyao Xia¹

¹ CIMS Research Center, Tongji University, Shanghai 201804, China
zhaojialee@tongji.edu.cn

² Engineering Research Center of Enterprise Digital Technology, Ministry of Education,
Shanghai 201804, China

Abstract. In the digital workshop with highly customized products, material delivery is the intermediate hub between warehouse management and production, which directly affects the products delivery speed and production efficiency. Due to the different kinds of materials, high flexibility of delivery plans and scattered workshop stations, the human-oriented material delivery mode still plays an important role. In this paper, augmented reality (AR) technology is introduced into the material delivery process to improve the level of system information visualization and human-machine interaction, reduce redundant labor. This paper also introduces a novel AR markerless 3D registration technology based on scene feature recognition and visual relocalization module. It helps to realize high-precision and robust 3D registration of virtual information in the whole workshop and basic material delivery path guidance function. This AR system has completed the transformation from image-level or object-level 3D registration to region-level 3D registration, which is helpful to broaden the scope of industrial application. Finally, the feasibility of the system is verified in a workshop production unit. All necessary information during material delivery process can be correctly and timely transmitted to workers with Microsoft HoloLens 2 (a high-performance AR headset). The AR assisted material delivery system effectively reduces the probability of missing delivery or wrong delivery and logistics cost.

Keywords: Material delivery · Augmented reality · Markerless 3D registration · Digital workshop · Path guidance · Human-machine interaction · HoloLens 2

1 Introduction

At present, the Internet of Things, big data, artificial intelligence, cloud computing, 5G, mobile Internet and other new-generation information technologies are advancing with each passing day. With the deep integration and application of manufacturing ontology technology and emerging technology, the Fourth Industrial Revolution, characterized by intelligence, has developed vigorously. The traditional manufacturing mode of centralized production has gradually changed to decentralized and socialized manufacturing mode. The manufacturing paradigm has changed from mass production to personalized, refined, single-piece, small-batch, multi-variety and customized production [1–3].

Workshop is the basic element and component of manufacturing industry. The construction of intelligent workshop is the premise and foundation of intelligent manufacturing. Therefore, the transformation of workshop production mode is one of the most direct manifestations of the transformation of manufacturing paradigm. Logistics management is the “third profit source” for manufacturing enterprises [4]. How to effectively reduce the cost of logistics, improve the level of logistics management and service efficiency, limit the waste of resources is of key significance to realize service-oriented, personalized, flexible, collaborative, transparent development in the context of new-generation intelligent manufacturing.

The optimization of logistics system includes the optimization of logistics resource allocation, material delivery process optimization, warehouse management optimization, system automation level improvement and many other aspects. The material delivery process, as the intermediate hub between warehousing and production and the core subsystem of workshop logistics, directly affects the products delivery speed, production efficiency and determines whether the production can operate steadily and continuously. Due to the trend of highly customized production in digital workshop, there are many variety, shapes, batches and quantities of materials, and the flexibility of delivery plan is high. Therefore the human-oriented material delivery mode still plays an irreplaceable role at this stage. However, most of the relevant information is still recorded in paper materials. The information visualization level is too low and it is difficult to form complete and effective data records. It often leads to wrong delivery and missing delivery, chaotic delivery process, secondary delivery and time-consuming growth, etc. In addition, due to the uneven distribution of warehouses and stations in the workshop, the material delivery path of the workers is relatively complex, and it is hard to choose the optimal path, resulting in the reduction of delivery efficiency. These above problems not only increase the cost of manpower, transportation, materials and time, but also lead to complicated delivery procedures, high work pressure and reduced enthusiasm of workers. It is unable to meet accurate, punctual, intelligent and other comprehensive indicators required by the new generation of logistics system.

In view of the above problems, this paper applies augmented reality (AR), a new human-machine interaction method, to the material delivery process of digital workshop. On the basis of the comprehensive interconnection of various production factors in the workshop, AR System obtains the relevant data in real time, and timely pushes the virtual information to the delivery workers with Head Mounted Display (HMD), so as to simplify the delivery process. It effectively improves the level of system information visualization, the integrity and traceability of data records, and reduces the loss of working hours and material resources. Further, in order to realize the seamless fusion of virtual information with physical space and the deep interaction between delivery workers and workshop, it is necessary to complete the high-precision, reliable and robust localization of virtual information in the physical workshop, that is, 3D registration function. This AR system realizes markerless 3D registration function based on three modules: dense 3D reconstruction, visual relocalization and fusion tracking. Compared with the current 3D registration methods, which rely on QR codes, pictures etc., it does not need to place any manual markers in the scene in advance. Thus there will be no failure of AR system due to falling off, shielding or damage of manual markers. Besides the aesthetics has

also been improved. Based on the idea of scene feature recognition & matching, this method can accurately register the virtual information at the regional level, and the scope of industrial application has been greatly widened. It does not depend on the specific targets in the scene but uses the overall scene feature. Therefore, even if the physical objects in the workshop change dynamically to a certain extent, the system can still achieve high-precision 3D registration and has good robustness.

The rest of the paper is organized as follows. In Sect. 2, research status of material delivery in the workshop, augmented reality technology and its industrial applications are summarized. Section 3 presents the basic application framework of augmented reality assisted material delivery mode in digital workshop and the principle of markerless 3D registration technology based on scene feature recognition & visual relocalization. In Sect. 4, the augmented reality system is preliminarily verified in a workshop production unit. Sect. 5 is conclusion and future work.

2 Literature Review

2.1 Research on Material Delivery Process

Material delivery is the core subsystem and connection hub of workshop logistics and the key step to realize intelligent logistics. In order to meet the complex material delivery needs of single-piece, small-batch, multi-type and customized products, a large number of scholars have explored the methods of logistics system optimization. On the basis of understanding the development status of an auto parts company and the workshop production mode, Zhu et al. [5] analyzed the causes of picking waste, inventory waste and material handling waste, and improved the material delivery mechanism from two aspects: picking process and the distribution of warehouses. Yu et al. [6] built a material delivery path optimization model with time window constraints in the workshop of electronic assembly manufacturing enterprises by analyzing the workshop layout, product process routes and materail delivery mode. The simulation shows that the optimization model can improve the delivery efficiency. Aiming at the problem of single-piece and small-batch material delivery process in the assembly of hydraulic pump in an enterprise, Chen et al. [7] took the shortest total delivery time as the objective function, built a workshop station centered delivery model and solved it by genetic algorithm (GA), and finally verified the feasibility in the digital workshop of the corresponding hydraulic component manufacturing enterprise. Huang et al. [8] introduced the idea of lean logistics into the material delivery process, built a smart three-dimensional warehouse, promoted the informatization process of the workshop system through advanced technologies such as RFID and Internet of things, and realized the functional requirements of delivering the right material to the right station at the right time. But at present, most of the improvement of material delivery process mainly focused on the optimization methods of the overall allocation of material delivery resources. There are still few examples of optimizing the delivery process from the perspective of material delivery workers.

2.2 Research on AR 3D Registration Technology and AR Application

Augmented Reality (AR) can effectively merge virtual and real scenes, enhance the scenes in the real world, and then display it to users through monitors, projectors, Head

Mounted Display (HMD) or other tools. It completes the real-time interaction between virtual and real world, and improves users ‘perception and information exchange ability. In recent years, AR has developed rapidly in basic theoretical research and application.

3D registration technology is the premise of achieving excellent virtual-real fusion effect. 3D registration can be summarized as: The system gets the pose of AR equipment according to real-time sensor data, updates the spatial coordinate system according to the user ‘s current perspective, and then accurately registers the virtual information in the real environment.

At present, the mainstream 3D registration technology in AR System includes three kinds: 3D registration based on hardware sensors, 3D registration based on computer vision and hybrid 3D registration technology [9]. 3D registration technology based on hardware sensors often GPS, inertial measurement unit (IMU), electromagnetic sensors, etc. When used alone, it often has different problems such as low registration accuracy or limited by application scenarios, etc. 3D registration technology based on computer vision comprehensively uses the relevant theories of computer vision & computer graphics to process the image obtained by the camera, calculates the relative pose of the virtual information in the environment and finally completes the registration process. It can be further divided into marker-based and markerless 3D registration methods. The marker-based 3D registration technology needs to place the manual markers in the real scene in advance. Due to the relatively perfect characteristics, low computational complexity, good real-time performance and high recognition accuracy of markers, it is popular in the early AR application, especially in the equipment with limited computing and processing capacity. However, once the markers fall off, are blocked or damaged, the AR system will not work normally, which limits the scope of industrial application to a certain extent. Figure 1 shows the AR application effect of the QR code-based 3D registration.



Fig. 1. QR code-based 3D registration for AR system

The markerless 3D registration technology does not need to place manual markers in advance, and the scene is more scalable. The realization idea is to complete the high-precision localization of virtual information in the real scene with the help of the existing geometric features (such as lines, surfaces and cylinders, etc.) in the environment, or the 3D model of the real object. This method can also be realized by simultaneous localization and mapping (SLAM). A single 3D registration technology is often difficult to meet the different requirements at the same time, such as computational efficiency, stability,

real-time performance, robustness, application scope, accuracy, etc. Hybrid 3D registration technology can achieve complementary advantages by integrating various methods. But the system development is difficult. In general, markerless 3D registration technology based on multi-sensor fusion is one of the main research directions of augmented reality technology in the future. The research focus is how to design an effective sensor fusion mechanism, improve the robustness of the system and reduce the computational complexity on the premise of ensuring the registration accuracy.

In terms of industrial AR application, Li et al. [10] realized dynamic display & update of pipe network data and basic inspection functions based on HoloLens. Zhao et al. [11] combined AR and wifi technology to achieve indoor localization and live navigation in the process of material sorting. Mourtzis et al. [12] used AR technology for fast retrieval and information visualization of storage goods. Xu et al. [13] combined AR with deep neural network for fault diagnosis of power equipment, which helps to improve the recognition accuracy. Zhu et al. [14] combined digital twin with augmented reality, and used augmented reality as a visualization tool of system data. Moreover, basic human-machine interaction mechanism was introduced in the manufacturing process to improve production efficiency, but the system did not involve deep intelligent control of the processing process. Aiming at the lack of effective integration of multi-source data and single human-machine interaction means, Liu et al. [15] proposed a multi-view interaction method based on augmented reality and established an information integration model for digital twin processing system. Fang et al. [16] improved the efficiency and accuracy of material sorting based on QR codes widely distributed in the workshop and Head Mounted Display (HMD). However, the system needs to post a large number of QR codes in various stations and shelves in the workshop, which is a tedious process and still has a large space for improvement.

2.3 Research Gap

Material delivery is one of the important process of production in digital workshop. Generally speaking, the shortcomings of material delivery process at this stage can be summarized as follows:

- (1) In the digital workshop with single-piece, small-batch, multi-variety and customized products, the relevant research of logistics system often focuses on the overall allocation and optimization of material delivery resources. There are few examples of optimizing the delivery process from the perspective of material delivery workers.
- (2) As a novel way of human-machine interaction, augmented reality pays attention to virtual-real interaction and fusion. With the help of 3D registration technology, the predefined virtual information generated by computer can be accurately added to the real scene. Through the application of augmented reality technology, the information visualization and human-machine interaction level of material delivery system can be greatly improved. However, at this stage, in order to realize the accurate 3D registration of virtual information, it is often necessary to place a large number of QR codes, pictures or other manual markers in the scene in advance. Once the markers fall off, are blocked or damaged, the AR system will not work normally,

which limits the scope of industrial application to a certain extent. There are still few reliable, robust, scene feature recognition based, region-level, markerless 3D registration methods for augmented reality system and related industrial application examples.

- (3) Due to the large number of material delivery items, high flexibility of delivery plans, complex storage areas and workshop stations, large activity space, it is often difficult for delivery workers to effectively choose the optimal delivery path, which leads to the increase of delivery time and redundant labor.

3 Augmented Reality Assisted Material Delivery System in Digital Workshop

3.1 System Framework

Human still plays an indispensable role in the context of new-generation intelligent manufacturing system. The improvement ideas of human-oriented material delivery in the intelligent logistics system can be summarized as follows:

Relying on manual paper recording of relevant delivery information is extremely inefficient and prone to mistakes. Due to the small amount of accommodated information and single means of human-machine interaction, traditional electronic production board limited to specific stations can not meet the functional requirements of complex logistics delivery scenes nowadays. Moreover, it cannot actively push useful information according to different situations. Therefore, it is necessary to select efficient, reasonable and intelligent visualization tools and information push methods.

With the support of the new generation of information technology, all production factors in the workshop are fully interconnected, and it is relatively easy to obtain the relevant data of production activities in the system. Through the augmented reality (AR) technology and Head Mounted Display (HMD), the system can dynamically and actively push virtual information generated by computer to delivery workers in real time. It supports a variety of interaction methods such as gesture, voice, virtual button and so on, which can greatly improve the shortcomings of information visualization in the previous material delivery process. In order to assist workers to delivery materials properly in the workshop with large work scope and uneven distribution of workshop stations, the markerless 3D registration method for AR System based on scene recognition and visual relocalization can be used to realize basic delivery path guidance function. In order to give full play to the subjective initiative of the delivery workers, they can selectively receive related system information, such as delivery plan, delivery items, localization information of workers, current time, estimated time of delivery, material type/quantity/supplier/batch/inventory, start/end stations and their distribution in the workshop, historical data and so on. In addition, they can also record and update the necessary data during the delivery process. This means that the system realizes dual-direction and effective human-machine interaction and data closed-loop, which provides necessary support for virtual and reality fusion, human-machine symbiosis in digital workshop.

To support above ideas and promote the integration of AR technology and existing manufacturing ontology technology, an optional system framework of Augmented Reality Assisted Material Delivery System in Digital Workshop is shown in Fig. 2. The system consists of Human-Cyber-Physical three spaces and Intelligent Logistics System Hub. Physical Space realizes comprehensive perception & interconnection of heterogeneous elements through ubiquitous sensors, edge computing devices and reliable data acquisition & transmission modules in the workshop. Virtual Workshop (Cyber space) realizes deep-seated mapping and control of physical space by constructing multi-level and multi-granularity digital model of the workshop. In human space, each delivery worker receives the corresponding plan and completes efficient material delivery with the help of AR technology. Intelligent Logistics System Hub completes the overall optimization of production and logistics resources based on production orders and performs iterative verification in the cyber space. Finally, it formulates the material delivery plan and pushes it to delivery worker in time. The markerless 3D registration module for Augmented Reality in Intelligent Logistics System Hub matches the scene feature with the reconstructed 3D model, so as to complete the high-precision localization of virtual information in the scene. This module is also the basis of AR assisted delivery path guidance function. In this novel material delivery mode, the perception, memory and execution ability of delivery workers are improved. The system is in the virtuous dual-direction cycle of “Human Assisting Machine” and “Machine Assisting Human”, which provides necessary support for realizing the hybrid and enhanced intelligence with human in the loop.

3.2 Material Delivery Process

As shown in the Fig. 3, the basic material delivery process in the digital workshop can be summarized as follows:

- (1) after receiving the new production order, the workshop system integrates the existing production factors and begins to formulate an appropriate production plan.
- (2) the iterative optimization and verification of production and logistics system are carried out by Intelligent Logistics System Hub in the virtual workshop.
- (3) the material delivery plan is generated and distributed to the delivery workers in the physical production workshop.
- (4) after receiving the delivery plan from the system server, the delivery workers with HoloLens 2 (a high-performance Head Mounted Display) is ready to delivery, and arrives at the target warehouses or stations with the help of augmented reality based path guidance module.
- (5) the delivery workers will place the specified type and quantity of materials into the material sorting & counting table of the corresponding station according to the system prompt, and completes the recording and updating of material delivery information.
- (6) repeat step (4) and (5) until the delivery plan is completed.

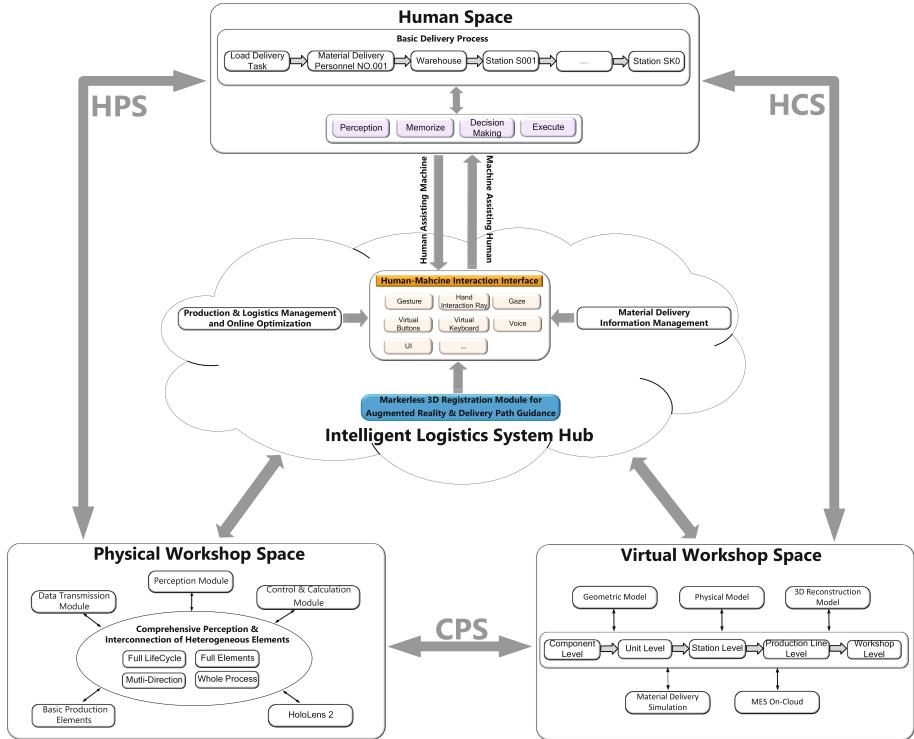


Fig. 2. System framework

3.3 Markerless 3D Registration Module for Augmented Reality System

Augmented reality (AR) system relies on high-precision 3D registration technology. However, in most industrial applications at this stage, a large number of manual markers need to be placed in advance in the real scene, which limits the AR application scope to a certain extent. In this paper, augmented reality markerless 3D registration technology is introduced into the material delivery process, and high-precision 3D registration can be realized only by the inherent scene feature in the workshop. The components of the markerless 3D registration module are shown in Fig. 4-1. The core components include three parts: 3D Reconstruction of Workshop, Regional Scene Recognition & Visual Relocalization and Fusion of Virtual and Real Information.

Among them, 3D reconstruction, as a common scientific research problem and core technology in the fields of computer vision, computer graphics, augmented reality, medical image processing, etc., restores the 3D information of objects through single view or multi view. In recent years, high-precision and realistic 3D reconstruction methods have attracted extensive attention. The dense 3D reconstruction method used in this paper is the continuation and improvement of KinectFusion [17] and BundleFusion [18]. The RGB-D SLAM (Simultaneous Localization and Mapping) can complete the processes of image acquisition, feature extraction, feature description, feature matching, camera trajectory and sparse scene information estimation, scene geometry and texture

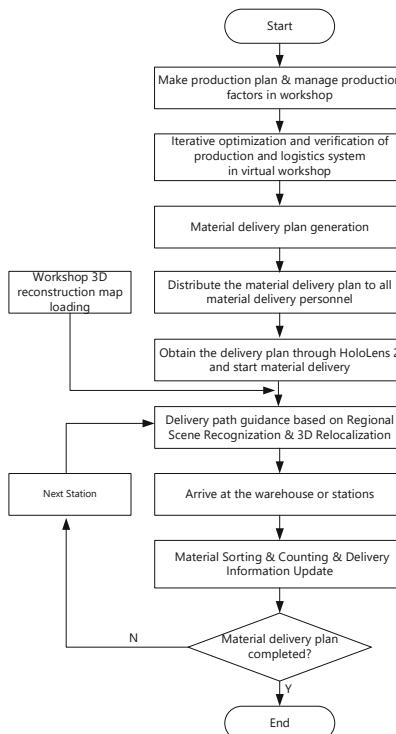


Fig. 3. Material delivery process

attribute restoration and so on. Compared with the traditional 3D modeling methods by professional software (such as 3dsmax, SolidWorks, CATIA, etc.), this method only needs RGB-D camera and simultaneous localization and mapping algorithm. It does not need a lot of manual intervention in the later stage and simplifies the workflow. It is worth mentioning that 3D reconstruction technology not only can be used as the support technology for AR system, but also provides a relatively novel solution for the digital description of large-scale scenes.

After the 3D reconstruction of the workshop is completed, the reconstructed model can be imported into Unity and other software development platforms, and the virtual information generated by computer can be predefined in the scene according to the actual needs. Then, in the material delivery process, real-time image data is matched with the scene feature of reconstructed model. After the matching is successful, the visual relocalization module obtains the relative pose between the virtual information and the actual scene. Finally, the fusion tracking module accurately adds the virtual information to the physical workshop according to the relative pose. The basic flow of visual relocalization is shown in the Fig. 4-2.

In addition, the localization module of HoloLens 2 (a high-performance Head Mounted Display) can be used to obtain the current localization information of the delivery workers. Combined with the markerless 3D registration technology and the optimized delivery path, it can realize AR assisted delivery path guidance function and further shorten the material delivery cycle.

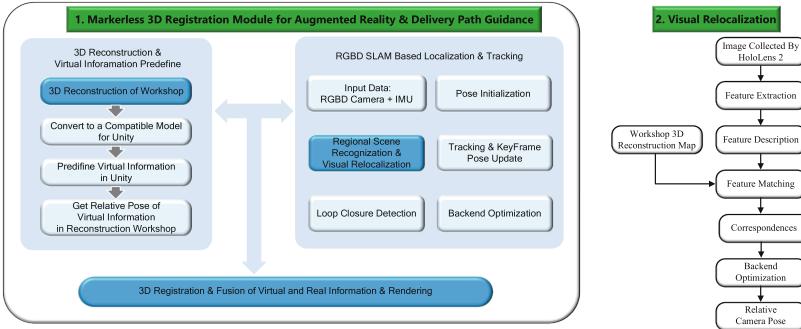


Fig. 4. Markerless 3D registration module for AR system

4 Case Study

4.1 Application Scenario

The intelligent production unit used for system function verification is shown in the Fig. 5. The physical space contains basic production factors, production equipment, communication equipment, logical control unit, etc. Production equipment mainly refers to two mechanical arms, laser engraving machine, rotary table, automatic lifting material table and two finished product shelves. iBox communication equipment completes the collection of data (such as real-time joint angle of manipulator, status information of laser engraving machine and warehouses, etc.), information exchange of cloud MES (Manufacturing Execution System) and transmission of relevant control commands. The production unit can complete the customization of the ring or bracelet patterns according to user requirements. The basic workflow is: after the user completes the online product customization, the production order will be arranged by cloud MES and the processing plan will be generated. Then the material delivery worker picks up the material from the warehouses and delivers them to the target workshop station. The production unit begins to process rough blanks and gives them to the finished product shelves after finishing processing. The digital model in cyber space is built based on Unity, including 3D action model and production optimization model. The composition and function of the system have been described in detail in [19].

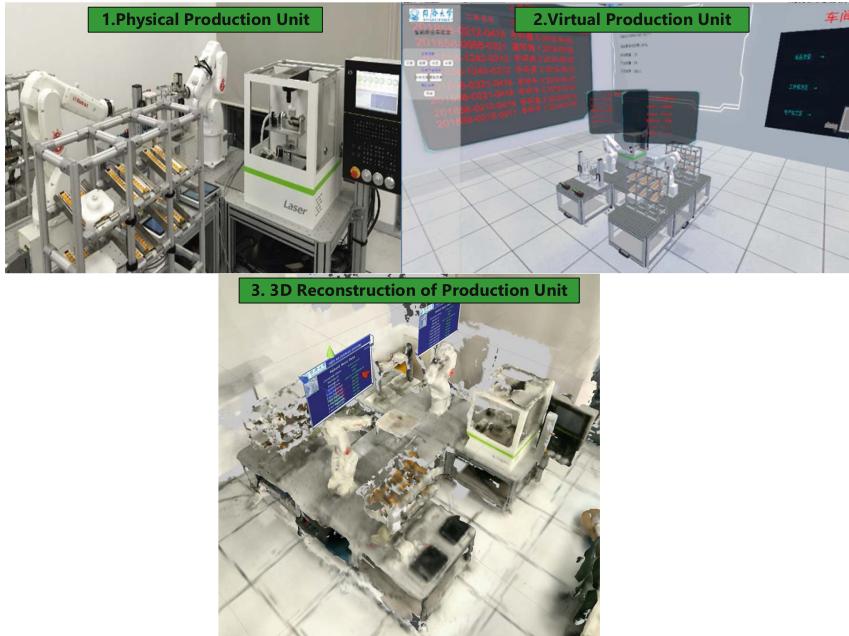


Fig. 5. Intelligent production unit

4.2 Augmented Reality Assisted Material Delivery System Development

Figure 5-3 shows 3D reconstruction model of the workshop based on RGB-D SLAM and the results of predefined and edited computer-generated virtual information on Unity development platform. Figure 6-1 shows the material delivery worker with HoloLens 2. Figure 6-2 to Fig. 6-4 show the relevant system information from the perspective of delivery worker. All data can be obtained from iBox edge computing device through Web Service. Display information includes real-time joint angle data of two mechanical arms, production equipment information, processing order information and production progress, program monitoring interface, material delivery plan & stations, different kinds of materials, etc. It can be seen that the 3D registration of virtual information in the workshop does not depend on the QR codes and other manual markers. The visual relocalization module actively pushes the virtual information to the user after successfully matching the scene features with the 3D reconstruction model of the workshop. Figure 6-5 shows the basic delivery path guidance function based on augmented reality. Figure 7 shows that even if the position of the mechanical arm in the scene changes greatly in the production and processing process, the virtual information can still achieve high-precision localization in the workshop. That is, the markerless 3D registration method adopted in this paper has certain robustness in the dynamic scene.



Fig. 6. Augmented reality system development



Fig. 7. Robustness verification of markerless 3D registration module

5 Conclusion and Future Work

In this paper, the human-oriented material delivery process in digital workshop with highly customized products is taken as the research starting point, and augmented reality (AR) technology is introduced. To sum up, the main contributions of this paper include the following three points:

- (1) This paper proposes an augmented reality assisted material delivery mode and its system application framework in digital workshop. This system uses HoloLens 2,

a high-performance and convenient HMD (Head Mounted Display), as the prototype system development platform and the medium of human-machine interaction. It breaks through the information barrier between human and machine and providing the necessary foundation for the seamless fusion of Human Space, Cyber Space and Physical Space. It is indeed an effective method to explore a new material delivery mode, and meets the requirements of constructing a perfect “Human-Cyber-Physical System” (HCPS) in the context of new-generation intelligent manufacturing.

- (2) In this paper, a high-precision and robust AR markerless 3D registration technology is applied to material delivery process to realize comprehensive display of workshop information and basic delivery path guidance function without placing a large number of manual markers (such as QR codes, pictures, etc.) in advance.
- (3) The feasibility of the proposed AR system was preliminarily verified in a highly customized workshop production unit. Workers with HoloLens 2 can timely obtain all necessary information during material delivery process. The system improves the perception, memory, executive ability of the delivery workers and the information visualization level. It also reduces the probability of missing delivery or wrong delivery and logistics cost.

The future work will mainly focus on the following aspects:

- (1) We will comprehensively consider the long-time, multi-stations, multi-tasks, highly flexible and complex production activities in the workshop and the simultaneous material delivery of multiple people.
- (2) This paper only realizes the basic material delivery path guidance function, and does not consider the situation that multiple workers carry out material delivery tasks at the same time and the dynamic change of delivery path. Therefore, the multi sensors (such as multi cameras, IMU and Ultra Wide Band, etc.) fusion localization methods and path planning algorithms can be introduced.
- (3) The form of human-machine interaction in this paper is still relatively simple. How to give full play to the overall control of virtual reality and partial advantages of augmented reality in material delivery process is still worth exploring [20].

Acknowledgement. This work is supported by National Key R&D Program of China (Grant No. 2017YFE0100900), Major scientific and technological innovation projects in Shandong Province (No. 2019TSLH0211) and Shanghai Science and Technology Innovation Action Plan (No. 19DZ1206800).

References

1. Zhou, J., Li, P., Zhou, Y., et al.: Toward new-generation intelligent manufacturing. *Engineering* **4**(4), 11–20 (2018)
2. Zhou, J., Zhou, Y., Wang, B., et al.: Human–cyber–physical systems (HCPSs) in the context of new-generation intelligent manufacturing. *Engineering* **5**(4), 624–636 (2019)

3. Zhou, J.: Intelligent manufacturing—main direction of “made in China 2025.” *China Mech. Eng.* **26**(17), 2273–2284 (2015)
4. The third profit source of enterprises—enter into the field of modern logistical management. *East China Econ. Manag.* (02), 46–47 (2001). <https://doi.org/10.19629/j.cnki.34-1014/f.2001.02.017>
5. Zhu, F.: Study on improvement of material distribution management of GT company production line. Dalian University of Technology (DUT) (2019). <https://doi.org/10.26991/d.cnki.gdllu.2019.004066>
6. Yu, W.T., et al.: Logistics distribution optimization of production workshop based on smart manufacturing. *Logist. Eng. Manag.* **42**(06), 6–11 (2020)
7. Chen, R.: Research on material distribution problem of digital components in hydraulic components. Hefei University of Technology (2019). <https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202001&filename=1019231957.nh>
8. Huang, H.: Research on intelligent storage and distribution in K company. Nanchang University (2020). <https://doi.org/10.27232/d.cnki.gnchu.2020.002308>
9. Han, Y., Li, T., Yang, D.: Overview of 3D tracking registration technology in augmented reality. *Comput. Eng. Appl.* **55**(21), 26–35 (2019)
10. Li, R.J.: Design and implementation of pipeline network patrol system based on mixed reality. University of Chinese Academy of Sciences (2020). <https://doi.org/10.27587/d.cnki.gksjs.2020.000016>
11. Zhao, J.J.: Application research of augmented reality technology in warehousing and picking. Shenyang University (2020). <https://doi.org/10.27692/d.cnki.gsydx.2020.000186>
12. Mourtzis, D., Samothrakis, V., Zogopoulos, V., et al.: Warehouse design and operation using augmented reality technology: a papermaking industry case study. *Procedia CIRP* **79**, 574–5794 (2019)
13. Xu, X., et al.: Research on power equipment fault identification method based on augmented reality technology. *Electron. Des. Eng.* **28**(23), 149–152+157 (2020). <https://doi.org/10.14022/j.issn1674-6236.2020.23.032>
14. Zhu, Z., Liu, C., Xu, X.: Visualisation of the digital twin data in manufacturing by using augmented reality. *Procedia CIRP* **81**, 898–903 (2019)
15. Liu, S., Lu, S., Li, J., et al.: Machining process-oriented monitoring method based on digital twin via augmented reality. *Int. J. Adv. Manuf. Technol.* **113**(11–12), 1–18 (2021)
16. Fang, W., et al.: Intelligent order picking method based on wearable augmented reality. *Comput. Integr. Manuf. Syst.* **27**(08), 2362–2370 (2021). <https://doi.org/10.13196/j.cims.2021.08.018>
17. Newcombe, R.A., Izadi, S., Hilliges, O., et al.: Kinectfusion: real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality, pp. 127–136. IEEE (2011)
18. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.* **36**(3), 76a (2017)
19. Xia, L., Lu, J., Zhang, H.: Research on construction method of digital twin workshop based on digital twin engine. In: 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE (2020)
20. Wang, X., et al.: Key technologies research and test of intelligent monitoring and control driven by AR/VR for fully mechanized coal-mining face. *J. China Coal Soc.* 1–17 (2022). <https://doi.org/10.13225/j.cnki.jccs.2021.1113>

Biomedical Informatics Theory and Methods



Safety and Efficacy of Short-Term vs. Standard Periods Dual Antiplatelet Therapy After New-Generation Drug-Eluting Stent Implantation: A Meta-analysis

Xiaohua Gao¹, Xiaodan Bi², Jinpeng Yang², and Meili Cheng^{1(✉)}

¹ Department of Cardiovascular, Heping Hospital Affiliated to Changzhi Medical College, Changzhi 046000, Shanxi, China

1320715155@qq.com

² Changzhi Medical College, Changzhi 046000, China

Abstract. To evaluate the safety and efficacy of short-term (6 months) and standard period (12 months) dual antiplatelet therapy (DAPT) after using a new generation of eluting drug stent (DES) in patients with coronary artery disease. Systematic review and meta-analysis were performed by searching PubMed, Embase, and Cochrane library databases for randomized controlled trials reporting on short-term (<6 months) and standard (12 months) DAPT after percutaneous coronary intervention implantation of a new-generation DES from inception to December 2020. The safety and effectiveness of treatment were compared. The primary endpoints were myocardial infarction, stent thrombosis, and major bleeding; secondary endpoints were all-cause mortality, cardiogenic death, any bleeding, stroke, target vessel revascularization, and net adverse clinical events. Sixteen studies (total 53,861 patients) were included. The incidence of major bleeding and any bleeding event in the short-term group was lower compared to standard-period group (major bleeding: 0.99% vs. 1.42%, RR = 0.68, 95%CI 0.57–0.82; major bleeding: 2.16% vs 3.19%, RR = 0.68, 95%CI 0.58–0.79; all P < 0.001), while there was no difference in myocardial infarction (2.16% vs 2.06%, RR = 1.04, 95%CI 0.93–1.17, P = 0.46), stent thrombosis (0.48% vs 0.38%, RR = 1.24, 95%CI 0.90–1.70, P = 0.20), all-cause mortality (1.70% vs 1.88%, RR = 0.90, 95%CI 0.80–1.02, P = 0.01), cardiogenic death (1.79% vs 0.97%, RR = 0.81, 95%CI 0.65–1.02, P = 0.08), stroke (0.72% vs 0.74%, RR = 0.96, 95%CI 0.78–1.19, P = 0.72), target vessel revascularization (3.87% vs 4.02%, RR = 0.99, 95%CI 0.87–1.12, P = 0.82), and net adverse clinical events (4.55% vs 5.01%, RR = 0.91, 95%CI 0.81–1.02, P = 0.09). Short-term DAPT significantly reduces the risk of bleeding compared with standard DAPT after percutaneous coronary intervention with a new-generation DES, without increasing the risk of mortality or ischemia.

Keywords: Percutaneous coronary intervention · Drug-eluting stent · Dual-antiplatelet therapy · Meta-analysis

1 Introduction

Coronary artery disease is the most common type of heart disease and a major cause of declining health status globally [24]. Dual antiplatelet therapy following the implantation of drug-eluting stents after the percutaneous coronary intervention is the standard therapy for the prevention of stent thrombosis in patients with coronary artery disease. However, the optimal duration of dual antiplatelet therapy (DAPT) after drug-eluting stent (DES) implantation is still not well defined. Some studies have shortened the duration of treatment to 3 or 6 months, while others have extended the duration of treatment beyond 12 months to determine the optimal duration of DAPT, maximize the anti-ischemic protection, and minimize bleeding. ACS combination for patients, type of drug stent, bleeding, and ischemic risks are important in determining the duration of DAPT [25]. Furthermore, a cost-benefit analysis of DAPT of different durations after percutaneous coronary intervention suggested that DAPT of 3 to 6 months is superior to DAPT of ≥ 12 months [26]. Also, forced interruption of DAPT due to poor adherence or bleeding increases the risk of adverse events (more common when the duration of DAPT is longer) [27]. Therefore, shortening the duration of DAPT may ease the global health burden.

Herein, we performed a meta-analysis to evaluate the safety and efficacy of short-term (≤ 6 months) and standard period (12 months) DAPT after using a new generation of DES.

2 Materials and Methods

2.1 Search Strategy

Randomized controlled studies reporting on short-term (≤ 6 months) and standard-period (12 months) DAPT following implantation of new-generation drug-eluting stents after percutaneous coronary interventions were searched in PubMed, Embase, and a Cochrane Library databases from inception to December 2020. The literature search principles were used to create a search formula with subject terms and free words. The search terms were: all relevant combinations of “coronary heart disease, CHD, acute coronary syndrome, ACS, myocardial infarction, MI, unstable angina, UA, stable angina, percutaneous coronary intervention, PCI, drug-eluting stent(s), DES, stent, platelet aggregation inhibitors, aspirin, prasugrel, ticagrelor, clopidogrel, P2Y12 inhibitor, dual antiplatelet therapy, DAPT, duration, death, mortality, cardiac mortality, stent thrombosis, bleeding, stroke, randomized controlled trials, random*, random allocation”. There was no language restrictions. PubMed has set up update alerts in order to include the latest studies.

2.2 Literature Inclusion/Exclusion Criteria

Inclusion criteria were: (1) study subjects: age ≥ 18 ; patients receiving DAPT following implantation of new-generation drug-eluting stents after percutaneous coronary interventions. New-generation DES was defined as any DES after the first-generation DES.

(2) Interventions: patients were divided into two groups according to the length of DAPT application in the study: short-term group ($DAFT \leq 6$ months) and standard-period group ($DAPT = 12$ months), with the experiment group being the short-term group and the control group being the standard-period group. (3) Endpoints: the primary endpoints were myocardial infarction (MI), stent thrombosis (ST), and major bleeding; the secondary endpoints were all-cause mortality, cardiovascular death, any bleeding, stroke, target vessel revascularization (TVR), net adverse clinical events (NACE). (4) Study type: all included studies were clinical randomized controlled experiments.

Exclusion criteria were: (1) non-randomized controlled experiments; (2) studies comparing short-term ($DAPT \leq 6$ months) long-term ($DAPT > 12$ months), studies comparing standard period ($DAPT = 12$ months), and long-term ($DAPT > 12$ months), or studies comparing short-term ($DAFT \leq 6$ months) and long-term ($DAPT > 12$ months); (3) literature with incomplete data and unavailable original data; (4) reviews, case reports, subgroup analyses, duplicate publications, and unpublished literature.

2.3 Literature Screening

Literature screening was performed independently by two researchers. Studies obtained from database search were imported into Note Express literature management software. Duplicate literature was then removed. Then, the two researchers performed initial screening by reading the titles and abstracts of literature to exclude irrelevant literature and re-screening by carefully reading the full text of the potentially included literature to determine the final inclusion or exclusion according to the inclusion and exclusion criteria. If the original data were not available or the data were incomplete, the authors were contacted. Disagreements were resolved by discussion.

2.4 Data Extraction

Data extraction followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [1]. For the finally included studies, the relevant information and data were extracted independently by the two researchers according to a pre-designed form. The main characteristics of the included studies comprised study name, sample size, study method, treatment protocol, follow-up time, and stent type. Baseline characteristics of patients included the proportion of patients with ACS or diabetes and mean age. The primary endpoints were MI, ST, and major bleeding; the secondary endpoints were all-cause mortality, cardiovascular death, any bleeding, stroke, TVR, and NACE. ST was defined according to the American Research Council (ARC) standard report [2] and included a combination of definite or probable ST observed and documented in studies. Major bleeding was also defined (Table 1). Definition of bleeding was based on the definition of the included studies or a combination of major and minor bleeding. NACE was defined as a combination of ischemic and bleeding events in studies. Disagreements were resolved by discussion.

2.5 Literature Quality Evaluation

The quality evaluation of the included literature was independently performed by two researchers. The quality assessment of the included RCTs was performed by applying

Table 1. Definition of major bleeding

Bleeding grade	Definition of major bleeding
BARC3/5 type	3a type: significant bleeding with a 3–5 g/dL decrease in hemoglobin; significant bleeding requiring blood transfusion
	3b type: significant bleeding with a ≥ 5 g/dL decrease in hemoglobin; pericardial tamponade; bleeding requiring surgical intervention or control (except dental, nasal, skin, and hemorrhoids); intravenous vasoactive drugs required
	3c type: intracranial hemorrhage (except cerebral micro-bleed, hemorrhagic transformation, including intraspinal hemorrhage); hemorrhage that impairs vision
	5 type: fatal hemorrhage
TIMI	1. intracranial hemorrhage (except cerebral micro-bleed ≤ 10 mm)
	2. Clinically visible hemorrhage with a decrease in hemoglobin concentration ≥ 5 g/dL
	3. Fatal hemorrhage (death within 7 days as a direct result of hemorrhage)
GUSTO	1. Hemodynamically impaired bleeding requiring intervention
	2. Intracranial hemorrhage
REPLACE-2	1. intracranial, intraocular, or retroperitoneal hemorrhage; clinically evident hemorrhage accompanied by a decrease in hemoglobin >3 g/L
	2. Any hemoglobin decrease of 4 g/L
	3. Transfusion of ≥ 1 U concentrated red blood cells/whole blood

the Cochrane collaboration's tool for assessing risk of bias [3]. The evaluation included seven aspects: generation of randomized sequences, allocation concealment, blinding of subjects and researchers, blinding of outcome evaluation, completeness of outcome data, selective reporting of study results, and other sources of bias. Each study was evaluated for the above evaluation components: low risk of bias, high risk of bias, or inconclusive. Disagreements arising during the literature quality evaluation were resolved by discussion.

2.6 Statistical Analysis

2.6.1 Statistical Indicators

The study applied RevMan 5.3 statistical software for statistical analysis and processing. A P value <0.05 was considered to be statistically significant. All outcome indicators in this study were dichotomous variables, and their combined effect indicators were expressed by Relative risk (RR) and 95% confidence intervals (CI); the M-H method (Mantel-Haenszel) was applied for statistical analysis.

2.6.2 Heterogeneity Test

Heterogeneity among studies was assessed by Q-test and I₂ statistics. The I₂ statistic indicates the proportion of the heterogeneous part among studies in the total variance of effect sizes [4], and larger values of the I₂ statistic suggest greater heterogeneity. Significant heterogeneity was considered if I₂ was >50% or P was <0.1. Considering the heterogeneity of included experiments and its potential impact on treatment outcome, we pre-specified a random-effects model for statistical analysis.

2.6.3 Publication Bias

If more than ten studies were included, studies were tested for publication bias by the Funnel plots method.

3 Results

3.1 Literature Search Results

Two thousand seven hundred forty-three publications were obtained by searching the database; 1,920 were obtained after eliminating duplicates, 1,862 irrelevant publications were excluded by reading the titles and abstracts, and 42 were further excluded by reading the full text of the remaining 58 publications for the following reasons: unavailability of primary sources (n = 3), subgroup analysis (n = 12), non-randomized controlled experiments (n = 5), reviews (n = 14), and protocol design discrepancies (n = 8). After screening, 16 publications were finally included (Fig. 1). In addition, 3 randomized controlled experiments on the duration of DAPT were not included because the minimum duration of treatment with DAPT was 12 months, i.e., these studies did not include short-term (≤ 6 months) DAPT protocols [5–7].

3.2 Characteristics of Included Studies/Baseline Characteristics of Patients

Sixteen randomized controlled experiments [8–23] including 53,861 patients were eventually included. Among the 16 studies, two had a short-term DAPT set at 1 month, six had a short-term DAPT set at 3 months, and the remaining eight had a short-term DAPT set at 6 months. Except for the DAPT-STEMI, GLOBAL LEADERS, REDUCE, and SMART-DATE studies, the clinical endpoints were all determined based on 12-month follow-up records. All included studies were published between 2012 and 2020. The main characteristics of all included studies are detailed in Table 2, and the baseline characteristics of patients are shown in Table 3.

3.3 Quality Evaluation of Included Studies

The risk of bias for each of the included randomized controlled experiments was listed according to the Cochrane Collaboration's tool for assessing the risk of bias (Fig. 2). The 16 studies all described the randomized sequence generation method. Although most of the experiments were open design studies, adverse events were defined using a

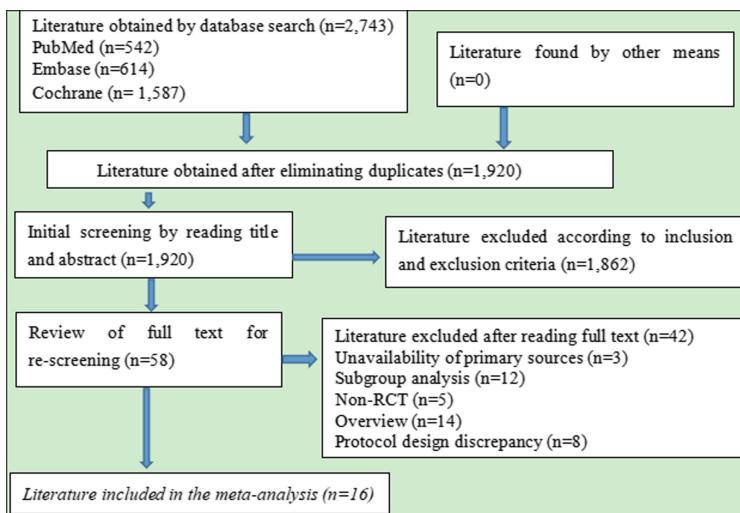


Fig. 1. Flow chart of literature screening

blind method so that the open design was not considered a source of significant bias. In addition, data from 15 clinical experiments were collected and analyzed by independent committees in each experiment, detecting a low risk of bias. There was no independent definition of clinical events in the GLOBAL LEADERS experiment. Missing data were reported in all experiments, and the reasons for missing data were described and analyzed by intention-to-treat analysis. All experiments were registered with clinical trials.gov and were certified as national clinical experiments.

3.4 Meta-analysis Results and Analysis of Publication Bias and Sensitivity

3.4.1 Primary Endpoints

3.4.1.1 Myocardial Infarction

The occurrence of myocardial infarction events was observed and recorded in 15 studies (53,370 patients). There was no statistical heterogeneity between the studies ($P = 0.60$, $I^2 = 0\%$). The incidence of myocardial infarction was 2.16% in the short-term group and 2.06% in the standard-period group, and the difference was not statistically significant (random-effects model: RR = 1.04, 95% CI 0.93–1.17, $P = 0.46$; Fig. 3). By visual estimation, both sides of the funnel plot were largely symmetrical, suggesting less publication bias in the included studies (Fig. 4).

3.4.1.2 Stent Thrombosis

Stent thrombosis events were observed and recorded in 14 studies (37,023 patients). There was no statistical heterogeneity between the studies ($P = 0.77$, $I^2 = 0\%$). The incidence of stent thrombosis was 0.48% in the short-term group and 0.38% in the standard-period group, and the difference was not statistically significant (random-effects model: RR = 1.24, 95% CI 0.90–1.70, $P = 0.20$; Fig. 5). By visual estimation, both sides of

Table 2. Study characteristics

Study	Year of publishing	First author	Country	Sample size	DAPT strategy	DAPT Course	Monotherapy after short-term DAPT	Follow-up time	Bleeding definition	Stent type
DAPT-STEMI	2018	Kedhi	4 countries	870	Aspirin+ Clopidogrel/ Prasugrel/ Ticagrelor	6 vs. 12	Aspirin	18	TIMI/ BARC	ZES
EXCELLENT	2012	Gwon	Korea	1,443	Aspirin+ Clopidogrel	6 vs. 12	Aspirin	12	TIMI	EES/SES
GLOBAL LEADERS	2018	Vranckx	18 countries	15,968	Aspirin+ Clopidogrel/ Ticagrelor	1 vs. 12	Ticagrelor	24	BARC	BES
I-LOVE-IT2	2016	Han	China	1,829	Aspirin+ Clopidogrel	6 vs. 12	Aspirin	18	BARC	BP-SES
ISAR-SAFE	2015	Schulz-Schupke	Germany	4,005	Aspirin+ Clopidogrel	6 vs. 12	Aspirin	15	TIMI	PES/SES/ EES/ZES/BES
IVUS-XPL	2016	Hong	North Korea	1,400	Aspirin+ Clopidogrel	6 vs. 12	Aspirin	12	TIMI	EES
OPTIMA-C	2018	Lee	Korea	1,368	Aspirin+ Clopidogrel	6 vs. 12	Aspirin	12	TIMI	BES/ZES
OPTIMIZE	2013	Feres	Brazil	3,119	Aspirin+ Clopidogrel	3 vs. 12	Clopidogrel	12	GUSTO/ BARC	ZES
REDUCE	2019	De Luca	Europe and Asia	1,496	Aspirin+ Clopidogrel/ Prasugrel/ Ticagrelor	3 vs. 12	Aspirin	24	BARC	COMBO

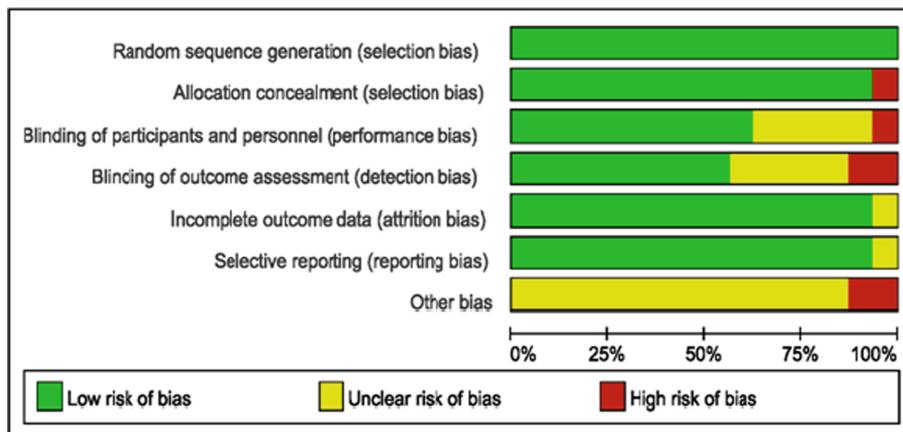
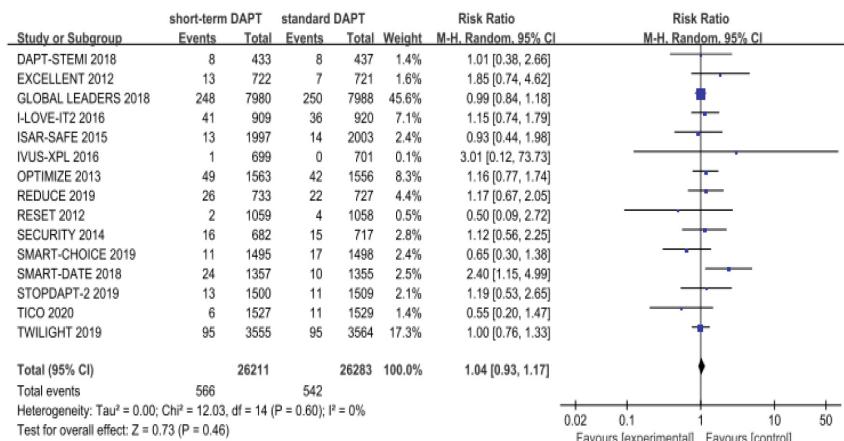
(continued)

Table 2. (continued)

Study	Year of publishing	First author	Country	Sample size	DAPT strategy	DAPT Course	Monotherapy after short-term DAPT	Follow-up time	Bleeding definition	Stent type
RESET	2012	Kim	North Korea	2,117	Aspirin+ Clopidogrel	3 vs. 12	Aspirin	12	TIMI	ZES/EES/SES
SECURITY	2014	Colombo	Europe	1,399	Aspirin+ Clopidogrel/ Prasugrel/ Ticagrelor	6 vs. 12	-	24	BARC	DES
SMART-CHOICE	2019	Hahn	Korea	2,993	Aspirin+ Clopidogrel/ Prasugrel/ Ticagrelor	3 vs. 12	Clopidogrel/ Prasugrel/ Ticagrelor	12	BARC	EES/SSES
SMART-DATE	2018	Hahn	Korea	2,712	Aspirin+ Clopidogrel/ Prasugrel/ Ticagrelor	6 vs. 12	Aspirin	18	BARC	EES/ZES/ BES/other
STOPDAPT-2	2019	Watanabe	Japan	3,009	Aspirin+ Clopidogrel/ Prasugrel	1 vs. 12	Clopidogrel	12	TIMI	EES
TICO	2020	Kim	Korea	3,056	Aspirin+ Clopidogrel/ Prasugrel/ Ticagrelor	3 vs. 12	Ticagrelor	12	TIMI	Bioresorbable polymer SES
TWILIGHT	2019	R. Mehran	11 countries	7,119	Aspirin+ Ticagrelor	3 vs. 12	Ticagrelor	12	TIMI/ BARC	DES

Table 3. Baseline patients characteristics

Study name	Short-term/ Standard- period DAPT (n)	Age (mean value)	Male (%)	Hypertension (%)	Diabetes (%)	Dyslipidemia (%)	Left anterior descending (%)	Left circumflex (%)	Right coronary (%)	ACS/Overall (n)
DAPT-STEMI	433/437	59.8/60.2	78/76	45/45	13/14	28/29	39/43	21/16	41/41	870/870
EXCELLENT	722/721	63/62.4	65.1/63.9	72.7/73.8	37.7/38.6	75.2/76.3	50.6/40.9	—	—	744/1443
GLOBAL LEADERS	7980/7988	64.5/64.6	76.6/76.9	74.0/73.3	25.7/24.9	69.3/70.0	41.2/42.0	24.3/24.5	31.6/30.7	7487/15968
I-LOVE-IT2	909/920	60.4/60	67.2/68.7	61.0/64.8	23.2/22.1	25.3/23.4	45.9/45.3	22.9/22.2	29.4/30.8	1496/1829
ISAR-SAFE	1997/2003	67.2/67.2	80.7/80.5	90.1/91.5	24.8/24.2	87.5/87.4	39.8/40.6	26.4/24	31.8/34	1601/4005
IVUS-XPL	699/701	63/64	67/70	63/65	36/37	68/65	55/56	20/18	25/26	686/1400
OPTIMA-C	683/684	62.8/64.4	70/67.8	62.4/63.9	29.1/29.7	29.9/28.5	57.7/51.6	19.6/24.1	22.7/24.3	692/1368
OPTIMIZE	1563/1556	61.3/61.9	63.5/63.1	86.4/88.2	35.4/35.3	63.2/63.7	47.9/46.6	23.4/24.3	27.6/27.7	1000/3119
REDUCE	751/745	61/60	82.6/77.3	50.7/50.7	21.6/19.5	46.3/44.9	48.0/44.2	—	—	1496/1496
RESET	1059/1058	62.4/62.4	64.4/62.9	62.3/61.4	29.8/28.8	57.7/59.9	52.7/53.6	21.0/19.2	26.3/27.1	601/2117
SECURITY	682/717	64.9/65.5	77.6/76.8	74.5/71.1	30.4/31.4	65.4/60.8	43/44	14.3/14.2	22/21.6	442/1399
SMART-CHOICE	1495/1498	64.6/64.6	72.7/74.2	61.6/61.3	38.2/36.8	45.1/45.5	48.8/50.4	21.6/19.9	28.3/27.8	1741/2993
SMART-DATE	1357/1355	62/62.2	74.9/75.9	49.9/48.7	26.9/28.1	24.2/25.2	56.6/61.0	24.4/25.1	37.2/36.2	2712/2712
STOPDAPT-2	1500/1509	68.1/69.1	78.9/76.5	73.7/74.0	39/38.0	74.4/74.8	55.2/56.6	17.9/20.2	29.1/27.2	1148/3009
TICO	1527/1529	61/61	79/80	50/51	27/27	61/60	48/48	19/19	30/31	3056/3056
TWILIGHT	3555/3564	65.2/65.1	76.2/76.1	72.6/72.2	37.1/36.5	60.7/60.2	—	—	—	4614/7119

**Fig. 2.** Quality evaluation of the included studies**Fig. 3.** Forest plot comparing the effect of DAPT on myocardial infarction in the short-term group and the standard-period group

the funnel plot were largely symmetrical, suggesting no significant publication bias in the included studies (Fig. 6).

3.4.1.3 Major Bleeding

Major bleeding events were reported in 16 studies (53,861 patients). There was no statistical heterogeneity between the studies ($P = 0.33$, $I^2 = 11\%$). The incidence of major bleeding was 0.99% in the short-term group and 1.42% in the standard-period group, suggesting that short-term DAPT significantly reduced the risk of major bleeding

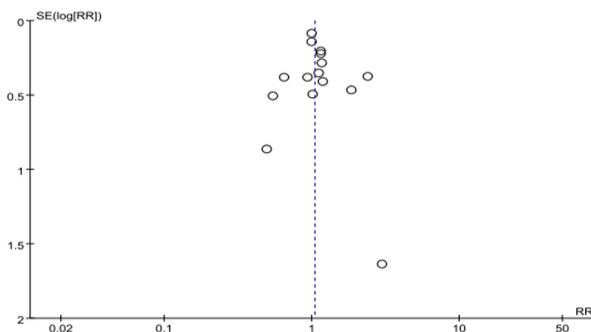


Fig. 4. Funnel plot comparing the effect of DAPT on myocardial infarction in the short-term group and the standard-period group

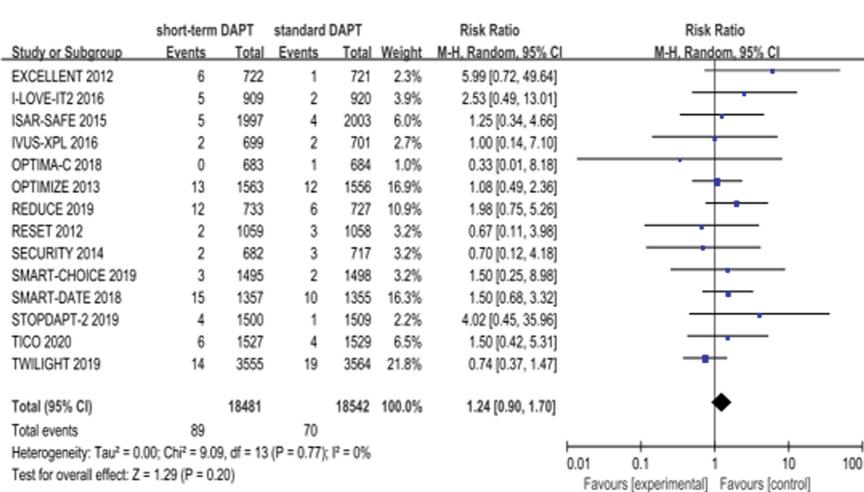


Fig. 5. Forest plot comparing the effect of DAPT on stent thrombosis in the short-term group and the standard-period group

compared with standard-period DAPT (random-effects model: RR = 0.68, 95% CI 0.57–0.82, $P < 0.0001$; Fig. 7). By visual estimation, both sides of the funnel plot were largely symmetrical, suggesting no significant publication bias in the included studies (Fig. 8).

3.4.2 Secondary Endpoints

3.4.2.1 All-Cause Mortality

All-cause mortality events were reported in 16 studies (53,861 patients). There was no statistical heterogeneity between the studies ($P = 0.74$, $I^2 = 0\%$). The incidence of all-cause mortality was 1.70% in the short-term group and 1.88% in the standard-period group, and the difference was not statistically significant (random-effects model: RR =

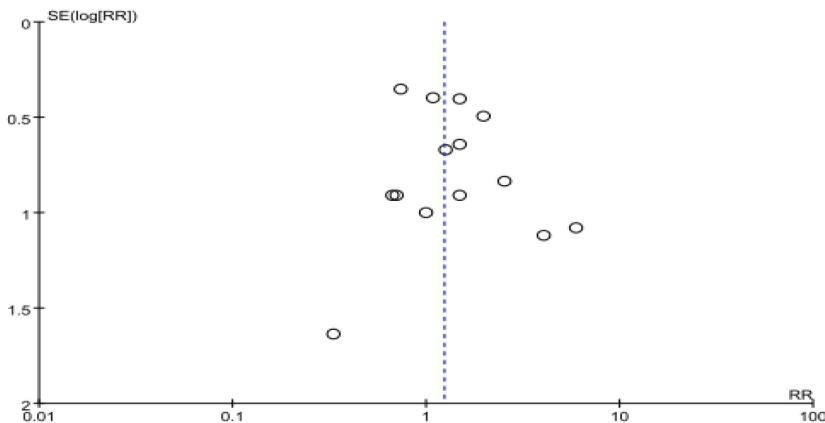


Fig. 6. Funnel plot comparing the effect of DAPT on stent thrombosis in the short-term group and the standard-period group

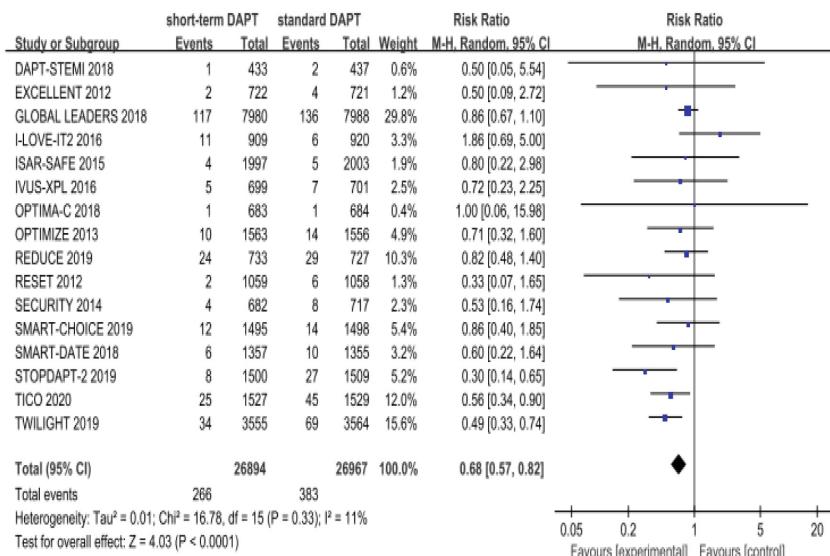


Fig. 7. Forest plot comparing the effect of DAPT on major bleeding in the short-term group and the standard-period group

0.90, 95% CI 0.80–1.02, $P = 0.11$; Fig. 9). By visual estimation, both sides of the funnel plot were less symmetrical, suggesting the possibility of publication bias in the included studies (Fig. 10).

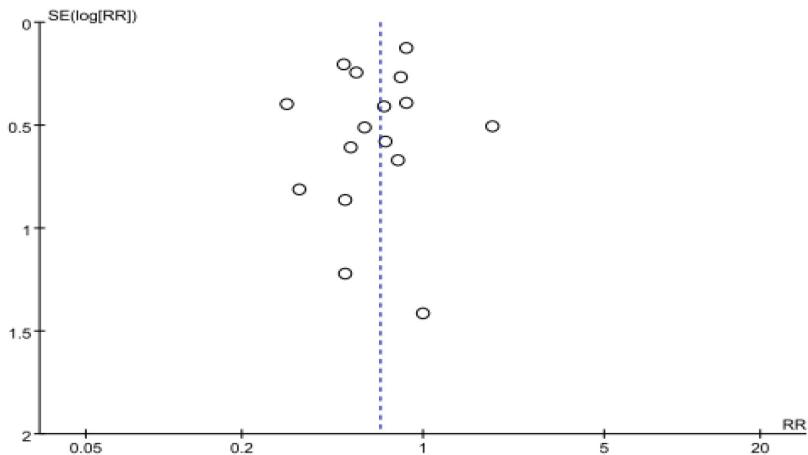


Fig. 8. Funnel plot comparing the effect of DAPT on major bleeding in the short-term group and the standard-period group

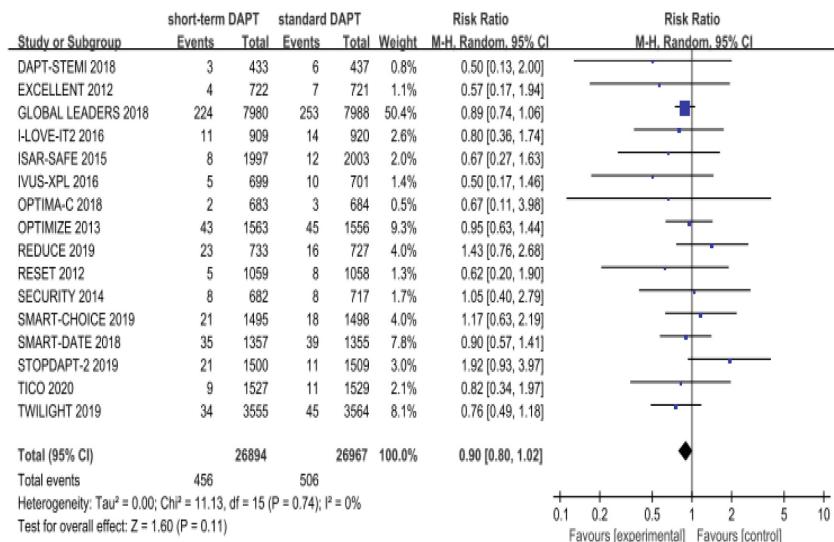


Fig. 9. Forest plot comparing the effect of DAPT on all-cause mortality in the short-term group and the standard-period group

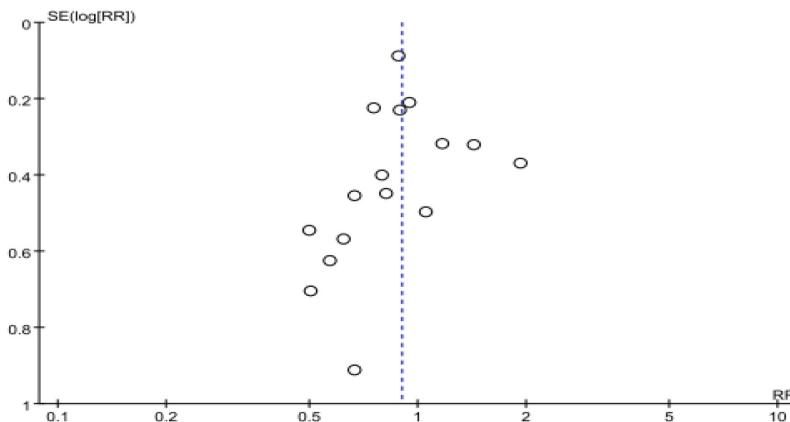


Fig. 10. Funnel plot comparing the effect of DAPT on all-cause mortality in the short-term group and the standard-period group

3.4.2.2 Cardiovascular Death

The occurrence of cardiovascular death events was observed and recorded in 14 studies (33,893 patients). There was no statistical heterogeneity between the studies ($P = 0.96$, $I^2 = 0\%$). The incidence of cardiovascular death was 0.79% in the short-term group and 0.97% in the standard-period group, and the difference was not statistically significant (random-effects model: $RR = 0.81$, 95% CI 0.65–1.02, $P = 0.08$; Fig. 11). By visual estimation, both sides of the funnel plot were slightly asymmetrical, suggesting no significant publication bias in the included studies (Fig. 12).

3.4.2.3 Any Bleeding

Any bleeding events were observed and recorded in 11 studies (24,998 patients). There was no statistical heterogeneity between the studies ($P = 0.96$, $I^2 = 0\%$). The incidence of any bleeding event was 2.16% in the short-term group and 3.19% in the standard-period group suggesting that short-term DAPT significantly reduced the risk of any bleeding event compared with the standard-period DAPT (random-effects model: $RR = 0.68$, 95% CI 0.58–0.79, $P < 0.00001$; Fig. 13). By visual estimation, both sides of the funnel plot were slightly asymmetrical, suggesting no significant publication bias in the included studies (Fig. 14).

3.4.2.4 Stroke

The occurrence of stroke events was observed and recorded in 15 studies (46,742 patients). There was no statistical heterogeneity between the studies ($P = 0.70$, $I^2 = 0\%$). The incidence of stroke was 0.72% in the short-term group and 0.74% in the standard-period group, and the difference was not statistically significant (random-effects model: $RR = 0.96$, 95% CI 0.78–1.19, $P = 0.72$; Fig. 15). By visual estimation, both sides of the funnel plot were largely symmetrical, suggesting no significant publication bias in the included studies (Fig. 16).

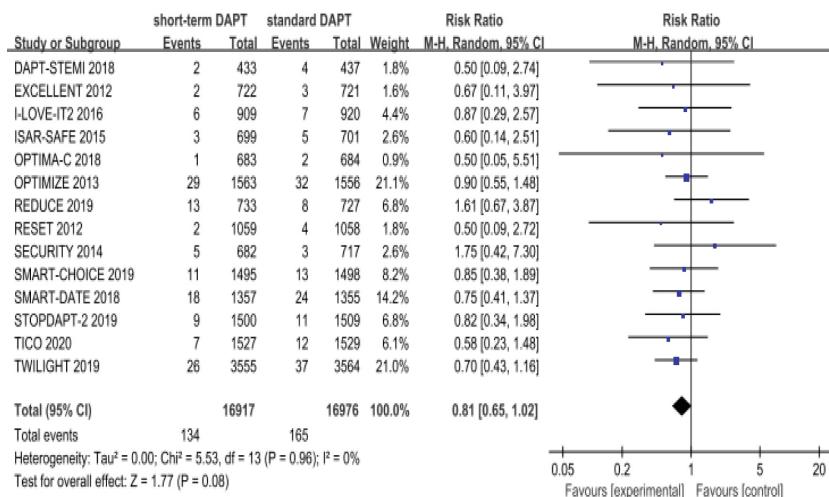


Fig. 11. Forest plot comparing the effect of DAPT on cardiac death in the short-term group and the standard-period group

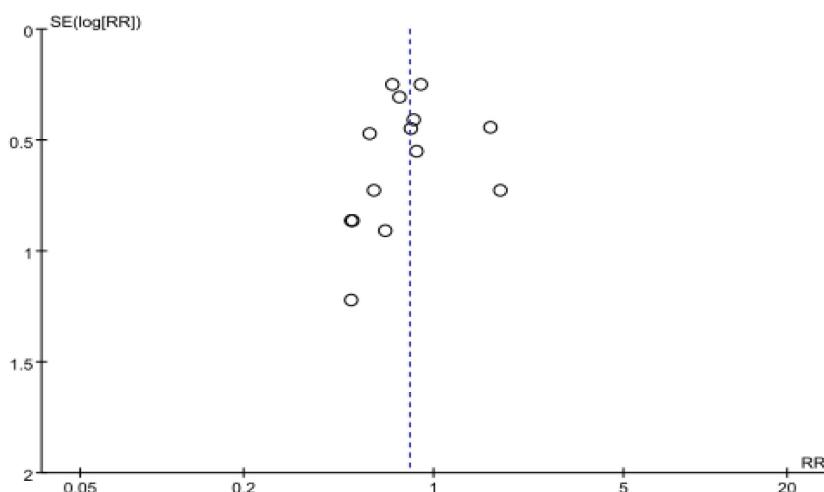


Fig. 12. Funnel plot comparing the effect of DAPT on cardiac death in the short-term group and the standard-period group

3.4.2.5 Target Vessel Revascularization

Target vessel revascularization events were observed and recorded in 9 studies (31,229 patients). There was no statistical heterogeneity between the studies ($P = 0.39$, $I^2 =$

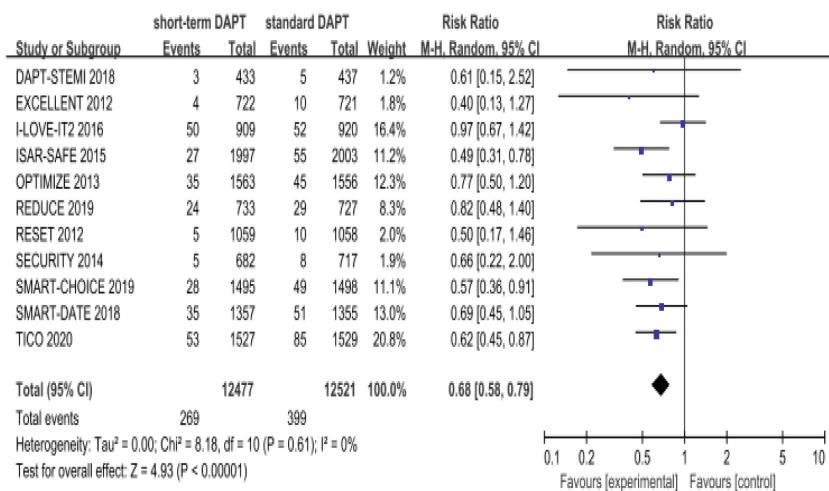


Fig. 13. Forest plot comparing the effect of DAPT on any bleeding in the short-term group and the standard group

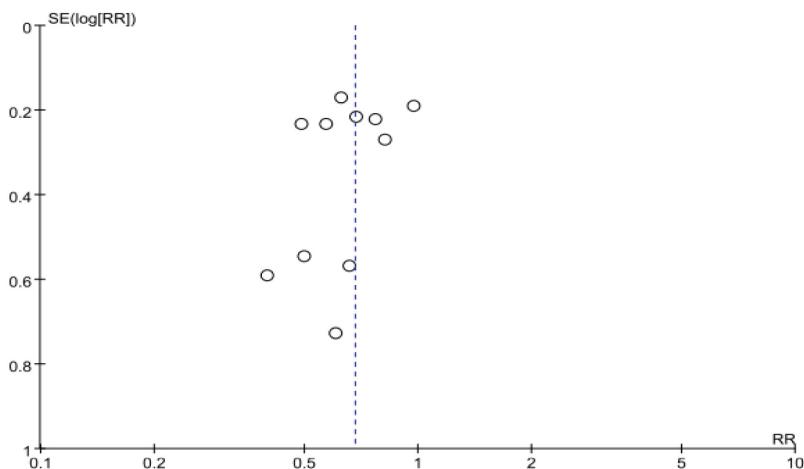


Fig. 14. Funnel plot comparing the effect of DAPT on any bleeding in the short-term group and the standard group

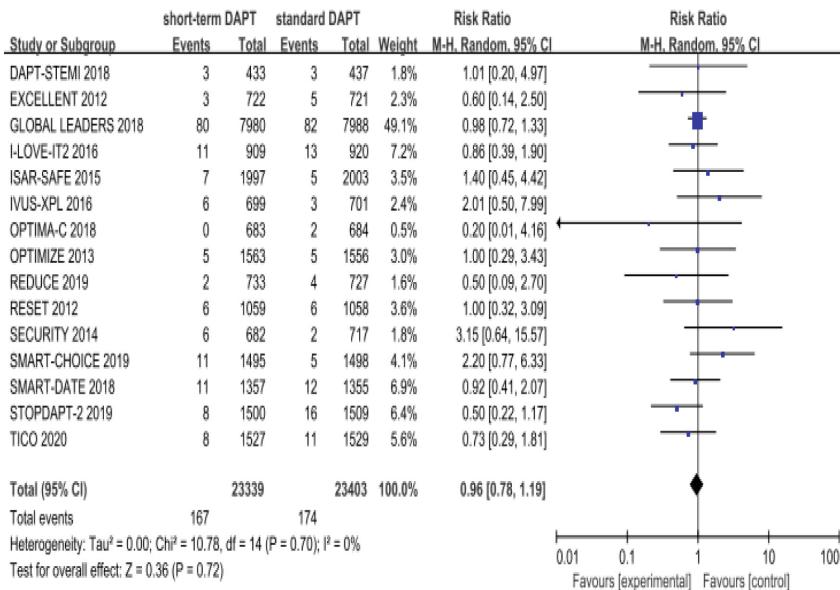
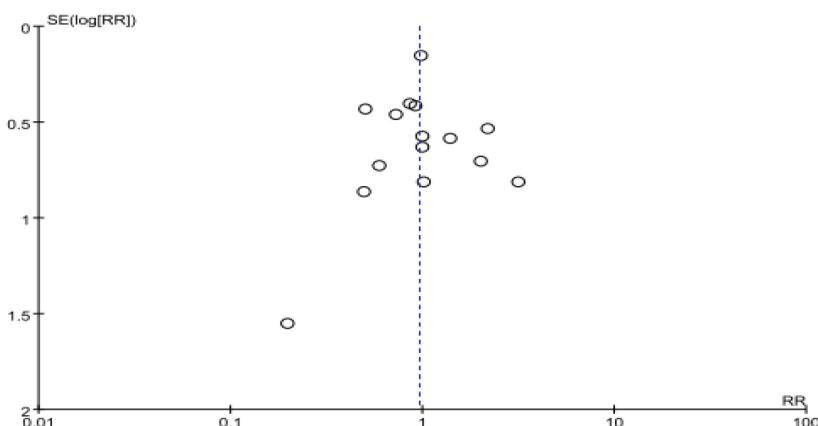


Fig. 15. Forest plot comparing the effect of DAPT on stroke in the short-term group and the standard-period group



6%). The rate of target vessel revascularization was 3.87% in the short-term group and 4.02% in the standard-period group, and the difference was not statistically significant (random-effects model: RR = 0.99, 95% CI 0.87–1.12, $P = 0.82$; Fig. 17). By visual estimation, both sides of the funnel plot were slightly asymmetrical, suggesting no significant publication bias in the included studies (Fig. 18).

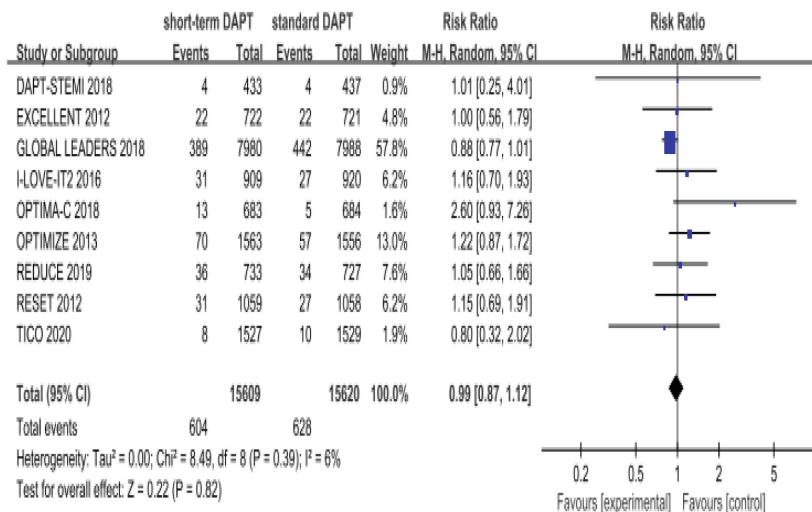


Fig. 17. Forest plot comparing the effect of DAPT on target vessel revascularization in the short-term group and the standard-period group

3.4.2.6 Net Adverse Clinical Events

Net adverse clinical events were observed and recorded in 12 studies (27,290 patients). There was no statistical heterogeneity between the studies ($P = 0.36$, $I^2 = 9\%$). The incidence of net adverse clinical events was 4.55% in the short-term group and 5.01% in the standard-period group, and the difference was not statistically significant (random-effects model: RR = 0.91, 95% CI 0.81–1.02, $P = 0.09$; Fig. 19). By visual estimation, both sides of the funnel plot were largely symmetrical, suggesting no significant publication bias in the included studies (Fig. 20).

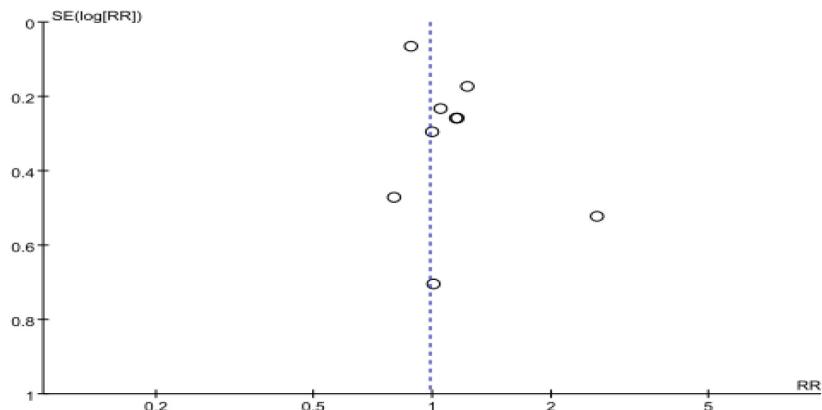


Fig. 18. Funnel plot comparing the effect of DAPT on target vessel revascularization

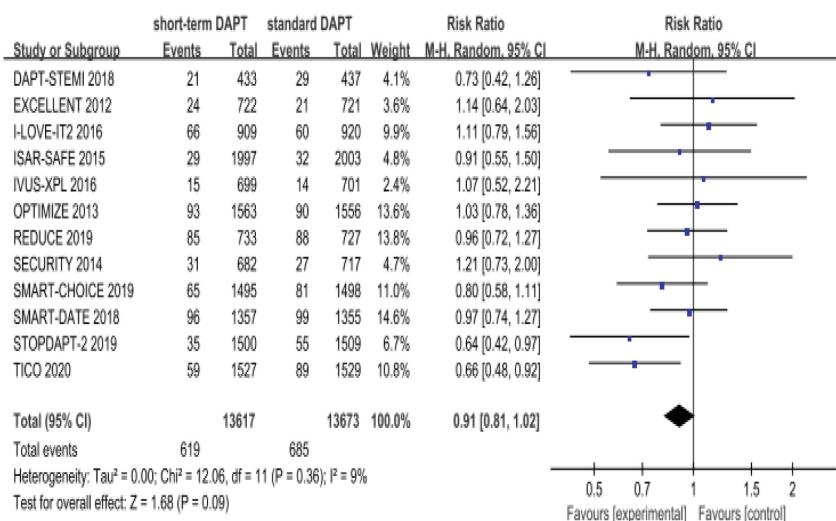


Fig. 19. Forest plot comparing the effect of DAPT on net adverse clinical events in the short-term group and the standard-period group

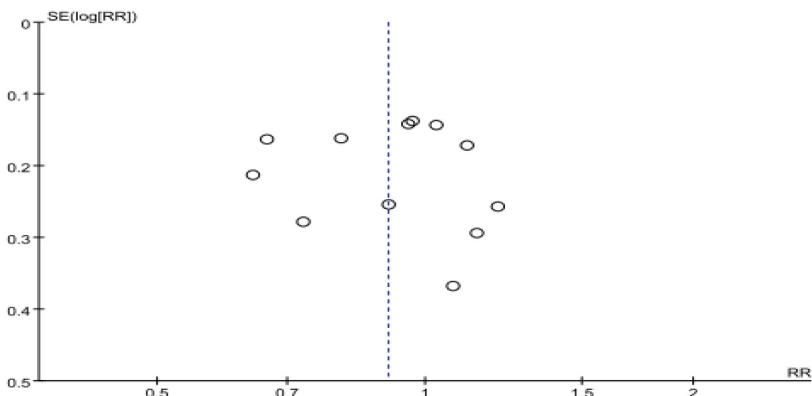


Fig. 20. Funnel plot comparing the effect of DAPT on net adverse clinical events in the short-term group and the standard-period group

4 Discussion

Growing evidence suggests that DAPT strategy depends on patient ischemia and bleeding risks and is affected by patient characteristics, clinical manifestations, comorbidities, co-medications, and procedural factors. Both ischemia and bleeding risks are strongly associated with adverse outcomes after percutaneous coronary intervention. For example, some studies suggested that major bleeding after the percutaneous coronary intervention is strongly associated with mortality, which may counteract the benefits of DAPT in anti-ischemic risk [28, 29]. Short-term DAPT may lower overall mortality by reducing the risk of major bleeding; thus, DAPT patients may benefit from a shorter duration of DAPT. However, despite a significant reduction in major bleeding in this meta-analysis, no statistically significant difference was found between the two treatment groups in all-cause mortality. This may be related to competing risks due to non-bleeding events. A recent meta-analysis based on 10 randomized controlled experiments concluded that the duration of DAPT beyond one year is associated with increased mortality, primarily due to an increased risk of non-cardiovascular death [30]. Another systematic review that included 11 randomized controlled experiments found no difference in all-cause mortality between 18 and 48 months of DAPT compared with 6 to 12 months of DAPT [31]. In addition, Mauri et al. found an increased risk of mortality in the extended DAPT treatment group, which was attributed to increased non-cardiovascular mortality due to cancer, bleeding, and trauma-related deaths [32]. Mehran et al. suggested an independent association between discontinuation of DAPT and stent thrombosis, particularly after implantation of first-generation DES [27]. Therefore, these data imply that short-term DAPT may increase the risk of ischemic events in patients undergoing percutaneous coronary intervention. However, with the optimized new-generation drug-eluting stent technology improving vessel healing and re-endothelialization, stent thrombosis

events have been significantly reduced [33, 34]. Recent randomized controlled studies suggested that short-term DAPT is not inferior to standard-period DAPT. In this meta-analysis, we extracted data on net adverse clinical events across studies, including the combined outcome of ischemic and bleeding events, and found no statistically significant difference between the two treatment protocols, which is consistent with the results of a net meta-analysis of individual patient data conducted by Palmerini et al. [35]. Moreover, Ndreppepa et al. performed a meta-analysis and showed similar rates of ischemia and bleeding between patients with DAPT \leq 6 months and those with DAPT = 12 months [36]. However, the higher incidence of major bleeding in standard-period DAPT compared with short-term DAPT was associated with the number of patients with acute coronary syndromes. This difference was more obvious when patients were only with acute coronary syndromes. The patients with acute coronary syndrome included in this meta-analysis (30,386 patients, representing more than 50% of the total sample size) showed good representativeness. As the duration of DAPT increases, the risk of bleeding increases; thus, clinicians should consider shortening the duration of DAPT in combination with effective medical management and risk factor modification to prevent ischemic events. In addition, non-cardiovascular mortality after prolonged DAPT was not counteracted by any benefit in reducing cardiovascular mortality, resulting in higher all-cause mortality. From this perspective, future directions need to identify the beneficiaries of shortened DAPT. Previous studies have attempted using DAPT to ASA monotherapy to reduce the risk of bleeding; for example, the dual antiplatelet therapy experiment was performed to compare the safety and efficacy of 12-month DAPT, 18-month ASA monotherapy, and 30-month DAPT [37]; within 3 months of shifting from DAPT to ASA monotherapy, the risk of bleeding was reduced, but the risk of stent thrombosis and myocardial infarction increased. In conclusion, because of the continued risk of bleeding with prolonged DAPT, clinicians should consider shortening the duration of DAPT in combination with effective medical management and risk factor modification to prevent stent thrombosis and myocardial infarction.

Currently, there are still no standard guidelines on the duration for patients treated with PCI. Since current guidelines are based primarily on evidence that was available prior to new-generation DES, only a small number of patients with acute coronary syndromes were included in recent experiments. The low endpoint event rates in many studies were insufficient to detect statistically significant differences, and some studies specifically excluded patients with a history of any bleeding or major comorbidities, making the bleeding and non-cardiovascular mortality lower in the included experiment subjects than in the general population. On the other hand, a shorter duration of DAPT may increase atherosclerotic thrombotic events in patients receiving drug therapy (e.g., for complex disease). Following the implantation of DES after percutaneous coronary intervention, short-term DAPT was not inferior to standard-period DAPT and it significantly reduced the risk of bleeding compared with standard-period DAPT, without statistically significant differences in the risk of myocardial infarction, stent thrombosis, all-cause mortality, cardiovascular death, stroke, target vessel revascularization, or net adverse clinical events. Thus, short-term DAPT may have a higher safety profile and similar efficacy in the general population with coronary artery disease compared with standard-period DAPT. Although short-term DAPT may be a feasible strategy for

patients after PCI, individualized treatment remains the preferred option without data from large experiments. To elucidate the impact of short-term DAPT on ischemia and bleeding risk, more relevant studies are needed to evaluate the application of short-term DAPT in contemporary clinical practice.

Compared with the previously published meta-analysis, this study has some advantages: (1) it covers the results of recent studies on short-term and standard-period DAPT following the implantation of drug-eluting stents after the percutaneous coronary intervention; (2) the included studies are more comprehensive and have larger sample sizes, which improves the statistical test efficacy, increases the credibility of the outcomes and makes the results more convincing; (3) the heterogeneity among the studies was low, and the results were more reliable.

This study also has a few limitations. The different follow-up times in the studies and the slightly different definitions of major bleeding and net adverse clinical events in the studies may reduce the accuracy, but no statistically significant heterogeneity was found in the outcome indicators in this study. Also, the number of patients with high ischemia or high bleeding risk in this study was insufficient. For example, the proportion of patients with ACS in the RESET and SECURITY experiments was 28.4% and 31.6%, respectively. In addition, patients with age >80, previous history of bleeding, oral anticoagulants, left main coronary artery diseases, or cardiogenic shock in some studies may affect assessing the risk of ischemia and bleeding.

To sum up, short-term DAPT significantly reduces the risk of bleeding, including major bleeding and any bleeding, compared with standard DAPT after percutaneous coronary intervention with the implantation of a new-generation DES, without increasing the risk of mortality or ischemia. Although short-term DAPT may be a viable strategy for patients after PCI, individualized treatment based on the risk of ischemia and bleeding is still preferred.

References

1. Liberati, A., Altman, D.G., Tetzlaff, J., et al.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**(jul21 1), b2700–b2700 (2009). <https://doi.org/10.1136/bmj.b2700>
2. Cutlip, D.E., Windecker, S., Mehran, R., et al.: Clinical end points in coronary stent trials: a case for standardized definitions. *Circulation* **115**(17), 2344–2351 (2007)
3. Higgins, J.P.T., Altman, D.G., Gotzsche, P.C., et al.: The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**(oct18 2), d5928–d5928 (2011). <https://doi.org/10.1136/bmj.d5928>
4. Higgins, J., Thompson, S.G., Deeks, J.J., et al.: Measuring inconsistency in meta-analyses. *BMJ* **327**(7414), 557–560 (2003)
5. Collet, J., Silvain, J., Barthélémy, O., et al.: Dual-antiplatelet treatment beyond 1 year after drug-eluting stent implantation (ARCTIC-Interruption): a randomised trial. *Lancet* **384**(9954), 1577–1585 (2014)
6. Physicians, A., O'Gara, P.T., Kushner, F.G., et al.: ACCF/AHA guideline for the management of ST-elevation myocardial infarction: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J. Am. College Cardiol.* **61**(4), e78–e140 (2013)

7. Whan, L.C., Jung-Min, A., Duk-Woo, P., et al.: Optimal duration of dual antiplatelet therapy after drug-eluting stent implantation: a randomized, controlled trial. *Circulation* **129**(3), 304–312 (2014)
8. Kedhi, E., Fabris, E., van der Ent, M., et al.: Six months versus 12 months dual antiplatelet therapy after drug-eluting stent implantation in ST-elevation myocardial infarction (DAPT-STEMI): randomised, multicentre, non-inferiority trial. *BMJ* **363**, k3793 (2018)
9. Gwon, H.C., Hahn, J.Y., Park, K.W., et al.: Six-month versus 12-month dual antiplatelet therapy after implantation of drug-eluting stents: the Efficacy of Xience/Promus Versus Cypher to Reduce Late Loss After Stenting (EXCELLENT) randomized, multicenter study. *Circulation* **125**(3), 505–513 (2012)
10. Pascal, V., Marco, V., Peter, J., et al.: Ticagrelor plus aspirin for 1 month, followed by ticagrelor monotherapy for 23 months vs aspirin plus clopidogrel or ticagrelor for 12 months, followed by aspirin monotherapy for 12 months after implantation of a drug-eluting stent: a multicentre, open-label, randomised superiority trial. *Lancet (London, England)* **392**(10151), 940–949 (2018)
11. Han, Y., Xu, B., Xu, K., et al.: Six Versus 12 months of dual antiplatelet therapy after implantation of biodegradable polymer sirolimus-eluting stent: randomized substudy of the I-LOVE-IT 2 Trial. *Circ. Cardiovasc. Interv.* **66**(2), B38–B38 (2015)
12. Schulz-Schupke, S., Byrne, R.A., Berg, J.M., et al.: ISAR-SAFE: a randomized, double-blind, placebo-controlled trial of 6 vs. 12 months of clopidogrel therapy after drug-eluting stenting. *Euro. Heart J.* **36**(20), 1252–1263 (2015). <https://doi.org/10.1093/eurheartj/ehu523>
13. Hong, S.J., Shin, D.H., Kim, J.S., et al.: 6-month versus 12-month dual-antiplatelet therapy following long everolimus-eluting stent implantation. *JACC: Cardiovas. Intervent.* **9**(14), 1438–1446 (2016). <https://doi.org/10.1016/j.jcin.2016.04.036>
14. Feres, F., Costa, R.A., Abizaid, A., et al.: Three vs twelve months of dual antiplatelet therapy after zotarolimus-eluting stents: the OPTIMIZE randomized trial. *JAMA* **310**(23), 2510–2522 (2013)
15. Luca, G.D., Damen, S.A., Camaro, C., et al.: Final results of the randomised evaluation of short-term dual antiplatelet therapy in patients with acute coronary syndrome treated with a new-generation stent (REDUCE trial). *EuroIntervent. J. EuroPCR Collabor. Work. Group Intervent. Cardiol. Euro. Soc. Cardiol.* **15**(11), e990–e998 (2019)
16. Byeong-Keuk, K., Myeong-Ki, H., Dong-Ho, S., et al.: A new strategy for discontinuation of dual antiplatelet therapy: the RESET Trial (REal Safety and Efficacy of 3-month dual antiplatelet Therapy following Endeavor zotarolimus-eluting stent implantation). *J. Am. College Cardiol.* **60**(15), 1340–1348 (2012)
17. Byoung-Kwon, L., Jung-Sun, K., Oh-Hyun, L., et al.: Safety of six-month dual antiplatelet therapy after second-generation drug-eluting stent implantation: OPTIMA-C Randomised Clinical Trial and OCT Substudy. *EuroIntervent. J. EuroPCR Collabor. Work. Group Intervent. Cardiol. Euro. Soc. Cardiol.* **13**(16), 1923–1930 (2018)
18. Yong, H.J., Bin, S.Y., Hyeon, O.J., et al.: Effect of P2Y12 inhibitor monotherapy vs dual antiplatelet therapy on cardiovascular events in patients undergoing percutaneous coronary intervention: the SMART-CHOICE randomized clinical trial. *JAMA* **321**(24), 2428–2437 (2019)
19. Hahn, J., Song, Y.B., Oh, J., et al.: 6-month versus 12-month or longer dual antiplatelet therapy after percutaneous coronary intervention in patients with acute coronary syndrome (SMART-DATE): a randomised, open-label, non-inferiority trial. *The Lancet* **391**(10127), 1274–1284 (2018)
20. Hirotoshi, W., Takenori, D., Takeshi, M., et al.: Effect of 1-month dual antiplatelet therapy followed by clopidogrel vs 12-month dual antiplatelet therapy on cardiovascular and bleeding events in patients receiving PCI: the STOPDAPT-2 randomized clinical trial. *JAMA* **321**(24), 2414–2427 (2019)

21. Kim, B., Hong, S., Cho, Y., et al.: Effect of ticagrelor monotherapy vs ticagrelor with aspirin on major bleeding and cardiovascular events in patients with acute coronary syndrome: the TICO randomized clinical trial. *JAMA* **323**(23), 2407–2416 (2020)
22. Mehran, R., Baber, U., Sharma, S.K., et al.: Ticagrelor with or without aspirin in high-risk patients after PCI. *N. Engl. J. Med.* **381**(21), 2032–2042 (2019)
23. Colombo, A., Chieffo, A., Frascheri, A., et al.: Second-generation drug-eluting stent implantation followed by 6- versus 12-month dual antiplatelet therapy: the SECURITY randomized clinical trial. *J. Am. College Cardiol.* **64**(20), 2086–2097 (2014)
24. Roth, G.A., Johnson, C., Abajobir, A., et al.: Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**(1), 1–25 (2017)
25. Costa, F., Klaveren, D.V., James, S., et al.: Derivation and validation of the predicting bleeding complications in patients undergoing stent implantation and subsequent dual antiplatelet therapy (PRECISE-DAPT) score: a pooled analysis of individual-patient datasets from clinical trials. *Lancet* **389**(10073), 1025–1034 (2017)
26. Arbel, Y., Bennell, M.C., Goodman, S.G., et al.: Cost-effectiveness of different durations of dual-antiplatelet use after percutaneous coronary intervention. *Can. J. Cardiol.* **34**(1), 31 (2018)
27. Mehran, R., Baber, U., Steg, P.G., et al.: Cessation of dual antiplatelet treatment and cardiac events after percutaneous coronary intervention (PARIS): 2 year results from a prospective observational study. *Lancet* **382**(9906), 1714–1722 (2013)
28. Rao, S.V., Grady, K.O., Pieper, K.S., et al.: A comparison of the clinical impact of bleeding measured by two different classifications among patients with acute coronary syndromes. *J. Am. Coll. Cardiol.* **47**(4), 809–816 (2006)
29. Budaj, A., Eikelboom, J.W., Mehta, S.R., et al.: Improving clinical outcomes by reducing bleeding in patients with non-ST-elevation acute coronary syndromes. *Eur. Heart J.* **30**(6), 655–661 (2008)
30. Palmerini, T., Benedetto, U., Bacchi-Reggiani, L., et al.: Mortality in patients treated with extended duration dual antiplatelet therapy after drug-eluting stent implantation: a pairwise and Bayesian network meta-analysis of randomised trials. *Lancet* **385**(9985), 2371–2382 (2015)
31. Bittl, J.A., Baber, U., Bradley, S.M., et al.: Duration of dual antiplatelet therapy: a systematic review for the 2016 ACC/AHA guideline focused update on duration of dual antiplatelet therapy in patients with coronary artery disease: a report of the American College of Cardiology/American Heart Assoc. *J. Am. College Cardiol.* **2016**, 1116–1139 (2016)
32. Mauri, L., Kereiakes, D.J., Yeh, R.W., et al.: Twelve or 30 months of dual antiplatelet therapy after drug-eluting stents. *South China J. Cardiol.* **371**(4), 2155–2166 (2014)
33. Dangas, G.D., Serruys, P.W., Kereiakes, D.J., et al.: Meta-analysis of everolimus-eluting versus paclitaxel-eluting stents in coronary artery disease. *JACC: Cardiovasc. Intervent.* **6**(9), 914–922 (2013)
34. De Luca, G., Smits, P., Hofma, S.H., et al.: Everolimus eluting stent vs first generation drug-eluting stent in primary angioplasty: a pooled patient-level meta-analysis of randomized trials. *Int. J. Cardiol.* **244**, 121–127 (2017)
35. Tullio, P., Diego, D.R., Umberto, B., et al.: Three, six, or twelve months of dual antiplatelet therapy after DES implantation in patients with or without acute coronary syndromes: an individual patient data pairwise and network meta-analysis of six randomized trials and 11 473 patients. *Eur. Heart J.* **14**, 1034–1043 (2017)

36. Misumida, N., Abo-Aly, M., Kim, S.M., et al.: Efficacy and safety of short-term dual antiplatelet therapy (≤ 6 months) after percutaneous coronary intervention for acute coronary syndrome: a systematic review and meta-analysis of randomized controlled trials. *Clin. Cardiol.* **41**(11), 1455–1462 (2018)
37. Mauri, L., Kereiakes, D.J., Yeh, R.W., et al.: Twelve or 30 months of dual antiplatelet therapy after drug-eluting stents. *South China J. Cardiol* **37**1, 2155–2166 (2014)



Automated Diagnosis of Vertebral Fractures Using Radiographs and Machine Learning

Li-Wei Cheng¹ , Hsin-Hung Chou² , Kuo-Yuan Huang³ , Chin-Chiang Hsieh⁴ , Po-Lun Chu⁵, and Sun-Yuan Hsieh⁶

¹ Institute of Medical Informatics, National Cheng Kung University, No. 1, University Road, Tainan 70101, Taiwan
clw00186@gmail.com

² Department of Computer Science and Information Engineering, National Chi Nan University, No. 1, University Road, Puli Township, Nantou County 54561, Taiwan
chouhh@nccnu.edu.tw

³ Department of Orthopedics, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan 701, Taiwan
hkyuan@mail.ncku.edu.tw

⁴ Department of Radiology, Tainan Hospital, Ministry of Health and Welfare, Tainan 700, Taiwan

⁵ Department of Computer Science and Information Engineering, National Cheng Kung University, No. 1, University Road, Tainan 70101, Taiwan
f84064014@mail.ncku.edu.tw

⁶ Department of Computer Science and Information Engineering, Institute of Medical Information, Institute of Manufacturing Information and Systems, Center for Innovative FinTech Business Models, and International Center for the Scientific Development of Shrimp Aquaculture, National Cheng Kung University, No. 1, University Road, Tainan 70101, Taiwan
hsiehsy@mail.ncku.edu.tw

Abstract. Objective: People often experience spinal fractures. The most common of these are thoracolumbar compression fractures and burst fractures. Burst fractures are usually unstable fractures, often accompanied by neurological symptoms, and thus require prompt and correct diagnosis, usually using computed tomography (CT) or magnetic resonance imaging (MRI). However, X-ray images are the cheapest and most convenient tool for predicting fracture morphological patterns. Therefore, we built a machine learning model architecture to detect and differentiate compression fractures from burst fractures using X-ray images and used CT or MRI to verify the diagnostic outcome. Methods: We used YOLO and ResUNet models to accurately segment vertebral bodies from X-ray images with 390 patients. Subsequently, we extracted features such as anterior, middle, and posterior height; height ratios; and the height ratios in relation to fractures and adjacent vertebral bodies from the segmented images. The model analyzed these features using a random forest approach to determine whether a vertebral body is normal, has a compression fracture or has a burst fracture. Results: The precision for identifying normal bodies, compression fractures, and burst fractures was 99%, 74%, and 94%, respectively. The segmentation and fracture detection

L.-W. Cheng and H.-H. Chou—Contributed equally to this work.

K.-Y. Huang and S.-H. Hsieh—Contributed equally to this work.

results outperformed those of related studies involving X-ray images. Conclusion: We believe that this study can assist in accurate clinical diagnosis, identification, and the differentiation of spine fractures; it may help emergency room physicians in clinical decision-making, thereby improving the quality of medical care.

Keywords: Thoracolumbar X-ray image · Compression fracture · Burst fracture · Vertebral body segmentation · Machine learning model

1 Introduction

People often experience severe trauma, such as a car accident or fall from a height. Furthermore, they may experience osteoporosis combined with minor trauma, and this can easily cause a spine fracture. Spinal fractures tend to occur at the thoracolumbar junction. Thoracolumbar fractures are classified using clinical classification systems. Some are classified according to fracture type and mechanism [4, 15], and some are classified according to the state of the anatomical structure and nerves [7, 8]. The simplest and most widely used classification method is the three-column theory proposed by Denis [4]. In this method, the spine is divided into three parts: the anterior column, the middle column, and the posterior column.

Compression and burst fractures are the most common vertebral fractures [22]. Burst fractures frequently occur in high-energy trauma and are most commonly associated with falls and traffic accidents and occurred in 10% of cases [1]. The thoracolumbar spine, located at the transitional zone between the thoracic rib cage and lumbar spine, is susceptible to injury and fracture due to increased motion associated with this region of the body. If a thoracolumbar fracture involves the anterior and middle columns of a vertebral body, it is categorized as a burst fracture. Burst fractures are usually unstable fractures, often accompanied by neurological symptoms and spinal instability, and thus require definitive diagnosis, possible immediate surgery and, in the case of suspected burst fractures, warrant confirmation via computed tomography (CT) or magnetic resonance imaging (MRI) [3]. However, a resident or emergency room physician may have difficulty in identifying thoracolumbar fractures and even differentiating between a burst and a compression fracture from X-ray images.

Although CT or MRI examinations are more accurate than X-ray images in predicting fracture morphological patterns, X-ray images are the cheapest and most convenient tool. The high radiation dose of CT and the long examination time of MRI are not suitable for extensive screening of spinal fractures, and the medical cost is much higher than that of X-ray imaging. Several machine learning studies have focused on the segmentation or diagnosis of vertebral fractures using X-ray images, CT or MRI [11, 13, 16–18, 24, 25].

However, all of these studies are focused either on X-ray image, or CT/MRI alone, no studies involved the connection between both data sets. To the best of our knowledge, no other study has been conducted using X-ray image with corresponding MRI and/or CT scans to train with and verify, respectively. In addition, there is still a lack of research implementing machine learning models to study the differences between thoracolumbar burst and compression fractures through X-ray images. Therefore, we built a machine learning model architecture to detect and differentiate compression fractures from burst

fractures using X-ray images and used CT or MRI to verify the diagnostic outcome. Different from these studies [13, 16–18, 24, 25], we combined several models and steps that solved different problem scenarios to improve accuracy. Furthermore, our study is the first study to differentiate between burst and compression fractures using X-rays.

2 Methods

In this section, we describe the materials and methods utilized herein. First, we introduce our model architecture. Then, we introduce our approach for preprocessing X-ray images. After data preprocessing, we used the YOLOv4 model [2], and the ResUNet model [5] for segmentation. Finally, we extracted the features from the segmentation results and analyzed these features to diagnose the vertebral bodies as healthy or having a compression fracture or burst fracture.

2.1 Model Architecture

Our approach involved combining several models and steps to solve various problems. To solve the problem of image contrast, we used adaptive histogram equalization in data preprocessing [11, 19]. Because the input images had different sizes, we used the YOLOv4 model [2] for preliminary segmentation. ResUNet [5] was used to accurately segment vertebral bodies. Finally, after the experiment, we used a random forest model to identify healthy vertebral bodies or thoracolumbar burst or compression fractures through the analysis of the features of segmented images. The overall model architecture is presented in Fig. 1.

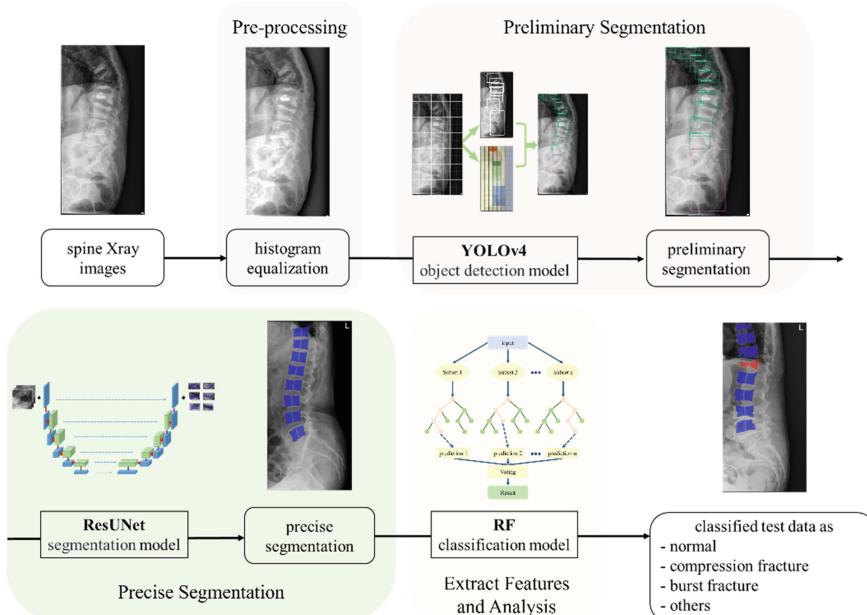


Fig. 1. Workflow of the proposed method

Adaptive histogram equalization [19] is a contrast enhancement method designed with broad applicability and demonstrable effectiveness. A machine learning network can be used to identify the location of each thoracolumbar vertebral body in input images of different sizes [11]. Therefore, we can use the YOLOv4 model [2], which is accurate and can be rapidly applied, to preliminary localize thoracolumbar vertebral bodies from T1 to L5. In the YOLO model [20], an entire image can be as the input of a neural network to directly predict the position of a bounding box, the confidence of a bounding box containing an object, and the category to which the object belongs.

U-net [21] is useful for biomedical image segmentation [10, 14, 23]. Resnet [9] is an outstanding image recognition model. ResUNet [5] is a combination of residual blocks and a U-net architecture. After obtaining preliminary segmented images, we used ResUNet to precisely segment vertebral bodies. The residual blocks, also known as ResBlocks, can avoid gradient diffusion. At this stage, a vertebral body with a screw or bone cement can be detected, as shown in Fig. 2. The cyan part is the vertebral body with a screw, and the deep blue part is that which had not been subjected to surgery.

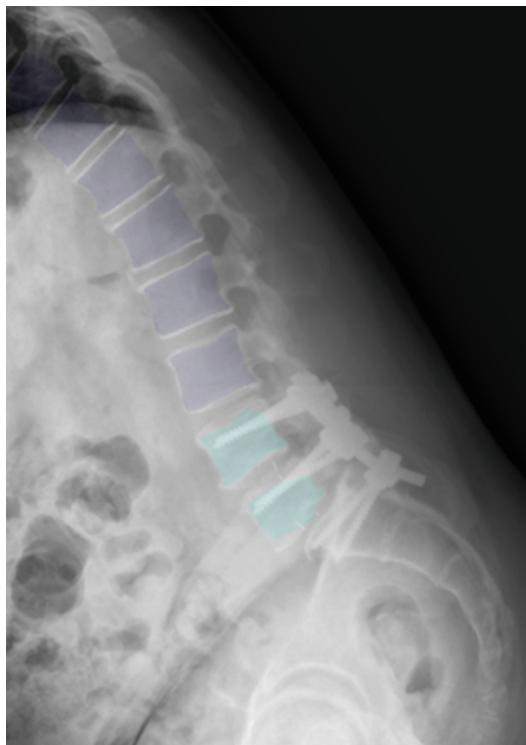


Fig. 2. The model identifies the vertebral body with the screw. (Color figure online)

2.2 Feature Analysis

Lateral view of thoracolumbar X-ray images was used to determine the type of vertebral fracture, and all was confirmed the diagnosis of burst or compression fractures by an orthopedic doctor. Such fractures can be diagnosed by the height and proportion of the anterior, middle, and posterior columns of the vertebral body. The ratios of the corresponding heights of the lower or upper vertebral bodies are also important considerations. We extracted features from the segmented images obtained using the previous steps. These data were then used to train machine learning models which can be used to identify the vertebral body type. We conducted experiments using the aforementioned models and selected the best one as our feature analysis model.

The main method of distinguishing between compression fractures and burst fractures involves determining whether there is damage to the middle column of the vertebral body. The distinction between compression and burst fractures is based on the height of the anterior and posterior vertebral bodies. If a vertebral body collapses, its anterior height should be discontinuous with the vertebral body above or below it.

The features we extracted were generally used for the diagnosis of vertebral fractures. First, the model measures the Ha , Hm , and Hp features, which respectively represent the anterior, middle, and posterior heights of vertebral bodies, as indicated in Fig. 3. Subsequently, we obtain R_{HmHa} , R_{HpHa} , R_{HmHp} , which denote the ratios of Hm to Ha , Hp to Ha , and Hm to Hp , respectively. These ratios in relation to upper or lower vertebral bodies are helpful for fracture diagnosis. The features extracted from the segmented images are shown in Table 1.

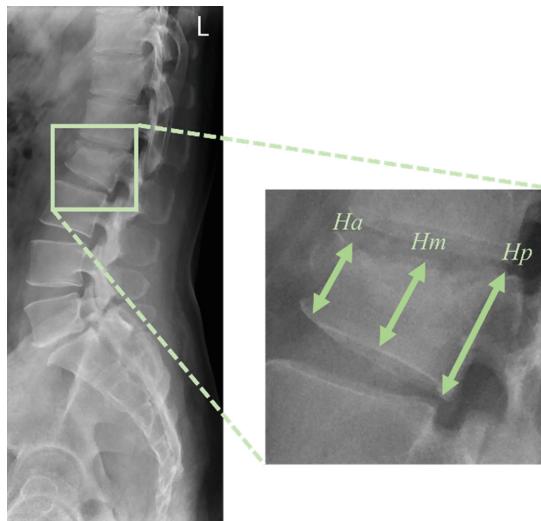


Fig. 3. Anterior, middle, and posterior height of a vertebral body

Table 1. Features extracted from the segmented images

Attribute	Description
Ha	Anterior height of the vertebral body
Hm	Middle height of the vertebral body
Hp	Posterior height of the vertebral body
R_{HmHa}	Ratio of Hm and Ha , $= \frac{Hm}{Ha}$
R_{HpHa}	Ratio of Hp and Ha , $= \frac{Hp}{Ha}$
R_{HmHp}	Ratio of Hm and Hp , $= \frac{Hm}{Hp}$
R_{Hi_lower}	$= \frac{Hi}{Hi \text{ of the adjacent lower vertebral body}}$, $Hi = Ha, Hm \text{ or } Hp$ (= -1 when there is no lower vertebral body)
R_{Hi_upper}	$= \frac{Hi}{Hi \text{ of the adjacent upper vertebral body}}$, $Hi = Ha, Hm \text{ or } Hp$ (= -1 when there is no upper vertebral body)
$R_{Hi_lower_encode}$	= 1 when $R_{Hi_lower} < 0.8$
$R_{Hi_upper_encode}$	= 1 when $R_{Hi_upper} < 0.8$

2.3 Data

In this study, we used data pertaining to orthopedics patients at our hospital and got human study approval from the IRB of our hospital. The data set contains 390 thoracolumbar X-ray images obtained from a picture archiving and communication system between January 2014 and December 2020. The range of the age is from 27 to 91. The average age of the patients is 75.12 ± 10.03 years. Of the patients, 302 are female and 88 are male. We were concerned about the presence of single-level or multi-level spinal fractures, which may have affected the experiment. The number of patients with a single-level spinal fracture, with a spinal fracture of levels 2 to 4, and with a spinal fracture more than 4-levels (≥ 5) was 271, 93, 26, respectively. The demographic data of the patients are presented in Table 2.

We labeled data in positions from T1 to L5 under four classes: normal, compression fracture, burst fracture, and others; the “others” class included a vertebral body containing a screw or bone cement. The labeled images were double-checked and evaluated with CT or MRI by an experienced orthopedic doctor. We used the VGG Image Annotator [6] to label the vertebral body. Examples of all four types of vertebral bodies are presented in Fig. 4. In addition to the four label classes, we used three level classes (upper thoracic level, thoracolumbar level, and lower lumbar level [24]) to present the position of the vertebral bodies. The total number of the vertebral bodies is 3634, and the details of them are shown in Table 3.

Table 2. Demographic data of the patients.

Patients with vertebral fracture	390
Age (years old)	75.12 ± 10.03
Gender - no. patients (%)	
Male	88 (22.6)
Female	302 (77.4)
Single/Multi-level spinal fractures - no. patients (%)	
Single level	271 (69.5)
2–4 level	93 (23.8)
5+ level	26 (6.7)

Table 3. The Data of the numbers of labeled vertebral bodies.

	Total	Normal	Compression	Burst	Others
Upper thoracic level (T1–T9)	729	634	38	41	16
Thoracolumbar level (T10–L2)	1809	1268	171	288	82
Lower lumbar level (L3–L5)	1096	979	16	3	98
Total	3634	2881	225	332	196

2.4 Implementation Details

Standard five-fold cross-validation was used for performance evaluations and comparisons [12], and we divided the data into five groups. We selected four groups as a training data set and one group as a test data set. After the testing, we calculated the average value as the model metrics. The advantage of this method is that all observations can be used for training and testing when only a small amount of data is available, and each observation is used for testing once.

In this experiment, Python 3.7, TensorFlow, and Keras were used to implement the machine learning framework and conduct data processing. We use OpenCV and VGG Image Annotator [6] as the image processing kit. All processes were performed on a machine with an AMD Ryzen 5 3600 CPU, 32 GB of DDR4 RAM, and NVIDIA GeForce RTX 2070 SUPER GPU.

The Dice coefficient was used to evaluate vertebral body segmentation quality. The Dice coefficient evaluates each vertebral body and then averages them. The Dice coefficient is defined as follows:

$$Dice = \frac{2 \times |P \cap G|}{|P| + |G|}$$

where P denotes the predicted segmentation results, G denotes the corresponding ground truth segmentation, and the operator | returns the number of labeled voxels.

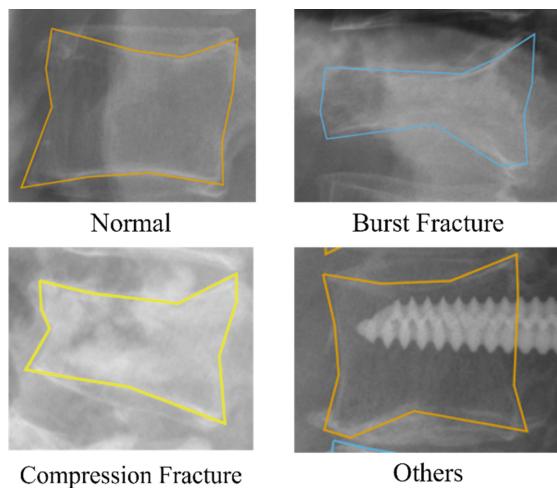


Fig. 4. Vertebral body labels

The accuracy, precision, recall and F1 score are used to evaluate diagnostic performance. We used a multi-class confusion matrix to present the diagnostic results. To verify the accuracy and feasibility of this modality, we used CT or MRI, and there is an example of a compression fracture and a burst fracture shown in Fig. 5.

3 Results

We used 390 vertebral X-ray images in the sagittal view, which include 3634 images of the vertebral body, to train and test our model with the five-fold cross-validation method. Segmentation results are shown in Fig. 6. The figures from left to right are the original image, the image which is labeled manually, the image which shows the segmentation results, and the one that presents overlapping labeled and segmented images. The average Dice coefficient for segmentation is 0.852.

We compared four feature analysis models, namely the support vector machine (SVM) model, random forest, multilayer perceptron, and k nearest neighbors. The comparison results are presented in Table 4. We used a multiclass confusion matrix to present our results, as shown in Fig. 7. We chose the random forest model for feature analysis after comparing the results. The precision for identifying normal bodies, compression fractures, and burst fractures was 99%, 74%, and 94%, respectively. If the model didn't distinguish between the compression fractures and burst fractures, the accuracy, precision, recall and F1-score will be 92.0%, 93.2%, 95.7% and 94.4%, respectively.

The time required to detect and identify a thoracolumbar burst or compression fracture by lateral radiograph analysis using our machine learning model, each step is completed in an average of 1 to 2 s after each image input. The entire process including graphic output can be completed within 30 s.

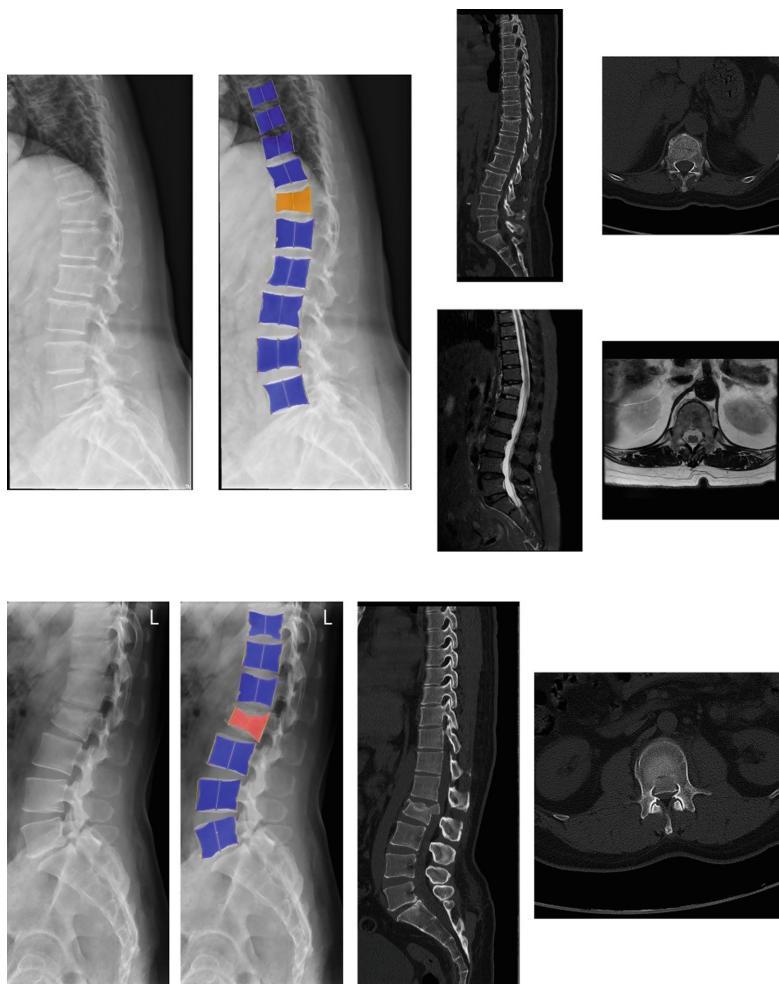


Fig. 5. CT or MRI used to verify the accuracy and feasibility of this modality.

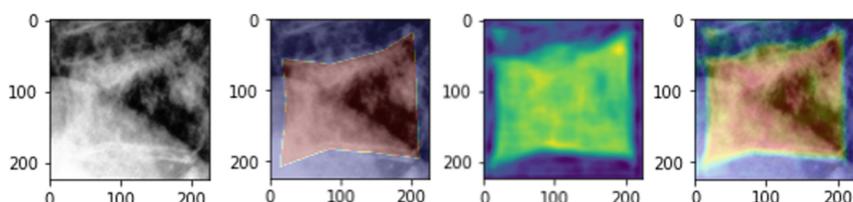
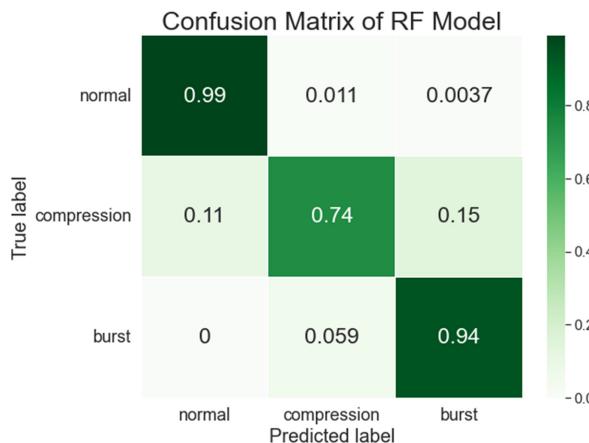


Fig. 6. Original, labeled, segmented, and overlapping images

Table 4. Model comparison related to thoracolumbar burst or compression fracture diagnosis

Model	Total	Normal			Compression			Burst		
	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
RF	0.98	0.94	0.99	0.96	0.77	0.74	0.75	0.86	0.94	0.90
KNN	0.89	0.90	0.98	0.94	0.75	0.71	0.73	0.93	0.87	0.90
SVM	0.89	0.91	0.98	0.94	0.60	0.50	0.55	0.86	0.70	0.78
MLP	0.89	0.89	0.99	0.94	0.83	0.45	0.59	0.91	0.69	0.78

**Fig. 7.** Confusion matrix of the random forest model

4 Discussion

By analyzing our diagnostic results, we revealed that some scenarios reduce the accuracy of differentiating between fracture types, including scoliosis and multilevel fracture. In multilevel fracture images, because the height of vertebral bodies adjacent to a fractured body is important for our study, continuous spine fractures render height ratio references invalid. The heights of the fracture combined with scoliosis obtained from the sagittal view of X-ray images will be inaccurate because of the presence of skewed vertebral bodies. Therefore, it is difficult to diagnose the pattern of vertebral fractures by lateral radiograph analysis alone.

Regarding the use of X-ray images, Kang Cheol Kim et al. [11] proposed an automatic X-ray image segmentation technique combining level-set methods and deep learning models such as pose-driven learning, which is used to selectively identify the five lumbar vertebrae, and M-net, which is employed to segment individual vertebrae. The performance of the proposed method was validated using clinical data; the center position detection error rate was $25.35 \pm 10.86\%$, and the mean Dice similarity measure was $91.60 \pm 2.22\%$. Kazuma Murata et al. [16] used a deep CNN to detect vertebral fractures on plain radiographs of the spine; their method achieved accuracy, sensitivity, and specificity of 86.0%, 84.7%, and 87.3%, respectively.

The differences between compression fractures and burst fractures are not sufficiently investigated in the literature. Therefore, we compared these fractures through vertebral fracture detection. A comparison of our vertebral fracture detection results with those of other studies is presented in Table 5.

Table 5. Vertebral fracture detection comparison with other studies

	Accuracy	Precision	Recall	F1-score
Kazuma Murata et al (for vertebral fracture using X-ray)	0.860	0.873	0.847	0.860
This study (for vertebral fracture using X-ray)	0.920	0.932	0.957	0.944

5 Conclusions

In this study, we proposed a method for diagnosing thoracolumbar vertebral burst or compression fractures on X-ray images. We used deep learning models such as YOLOv4 and ResUNet to accurately segment vertebral bodies in X-ray images. Subsequently, we extracted features from segmented images to analyze them using a random forest model. A combination of models for different situations improves our accuracy. The average Dice coefficient for the segmentation results was 85.2%. The precision for identifying normal bodies, compression fractures, and burst fractures was 99%, 74%, and 94%, respectively. The segmentation and fracture detection results outperformed those of related studies involving X-ray images.

It is recommended to arrange CT, MRI imaging examinations for patients highly suspected as burst fractures through our machine learning model to determine the definite diagnosis and evaluate whether surgical treatment is indicated. Besides, first aid with conservative treatment might be suggested for patients screened as compression fractures to reduce unnecessary medical waste. However, some unpredictable conditions may still occur and need to be further considered, such as the mechanism and energy of the injury and the actual clinical manifestations of the patient, such as whether there is neurological deficit, severe soft tissues damage, etc.

Our proposed machine learning method is easy to use and had good performance when used with the aforementioned data set from our hospital and will be improved to deal with diverse data sets and situations. This process is completed in a short time and can help to diagnose faster. We believe that this study can assist in accurate clinical diagnosis, identification, and differentiation of spine fractures; it can help emergency room physicians make accurate clinical decisions, thereby improving the quality of medical care.

References

1. Bensch, F.V., Koivikko, M.P., Kiuru, M.J., Koskinen, S.K.: The incidence and distribution of burst fractures. *Emerg. Radiol.* **12**(3), 124–129 (2006)

2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)
3. Dai, L.Y., Jiang, S.D., Wang, X.Y., Jiang, L.S.: A review of the management of thoracolumbar burst fractures. *Surg. Neurol.* **67**(3), 221–231 (2007)
4. Denis, F.: The three column spine and its significance in the classification of acute thoracolumbar spinal injuries. *Spine* **8**(8), 817–831 (1983)
5. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote. Sens.* **162**, 94–114 (2020)
6. Dutta, A., Zisserman, A.: The via annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2276–2279 (2019)
7. Groen, G.J., Baljet, B., Drukker, J.: Nerves and nerve plexuses of the human vertebral column. *Am. J. Anatomy* **188**(3), 282–296 (1990)
8. Haussler, K.K.: Anatomy of the thoracolumbar vertebral region. *Veterinary Clin. North Am. Equine Pract.* **15**(1), 13–26 (1999)
9. 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)
10. Ker, J., Wang, L., Rao, J., Lim, T.: Deep learning applications in medical image analysis. *IEEE Access* **6**, 9375–9389 (2017)
11. Kim, K.C., Cho, H.C., Jang, T.J., Choi, J.M., Seo, J.K.: Automatic detection and segmentation of lumbar vertebrae from x-ray images for compression fracture evaluation. *Comput. Meth. Program. Biomed.* **200**, 105833 (2021)
12. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, vol. 14, pp. 1137–1145. Montreal, Canada (1995)
13. Li, T., Wei, B., Cong, J., Li, X., Li, S.: S3egANet: 3D spinal structures segmentation via adversarial nets. *IEEE Access* **8**, 1892–1901 (2019)
14. Litjens, G., et al.: A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017)
15. Magerl, F., Aebi, M., Gertzbein, S., Harms, J., Nazarian, S.: A comprehensive classification of thoracic and lumbar injuries. *Eur. Spine J.* **3**(4), 184–201 (1994)
16. Murata, K., et al.: Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Sci. Rep.* **10**(1), 1–8 (2020)
17. Nicolaes, J., et al.: Detection of vertebral fractures in CT using 3D convolutional neural networks. In: Cai, Y., Wang, L., Audette, M., Zheng, G., Li, S. (eds.) CSI 2019. LNCS, vol. 11963, pp. 3–14. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39752-4_1
18. Pisov, M., et al.: Keypoints localization for joint vertebra detection and fracture severity quantification. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 723–732. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_70
19. Pizer, S.M., et al.: Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**(3), 355–368 (1987)
20. 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)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Savage, J.W., Schroeder, G.D., Anderson, P.A.: Vertebroplasty and kyphoplasty for the treatment of osteoporotic vertebral compression fractures. *JAAOS J. Am. Acad. Orthopaedic Surg.* **22**(10), 653–664 (2014)

23. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017)
24. Yabu, A., et al.: Using artificial intelligence to diagnose fresh osteoporotic vertebral fractures on magnetic resonance images. *Spine J.* **21**(10), 1652–1658 (2021). <https://doi.org/10.1016/j.spinee.2021.03.006>
25. Yousefi, H., Salehi, E., Sheyjani, O.S., Ghanaatti, H.: Lumbar spine vertebral compression fracture case diagnosis using machine learning methods on CT images. In: 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 179–184. IEEE (2019)



Cost and Care Insight: An Interactive and Scalable Hierarchical Learning System for Identifying Cost Saving Opportunities

Yuan Zhang², David Koepke³, Bibo Hao², Jing Mei², Xu Min², Rachna Gupta³, Rajashree Joshi³, Fiona McNaughton³, Zhan-Heng Chen⁴, Bo-Wei Zhao¹, Lun Hu¹, and Pengwei Hu¹ (✉)

¹ Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, China

hupengwei@hotmail.com

² IBM Research, Beijing, China

³ IBM Watson Health, Cambridge, USA

⁴ Shenzhen University, Shenzhen, China

Abstract. There has been an emerging interest in managing healthcare cost in the time of value-based care. However, many challenges arise in analyzing high-dimensional healthcare operational data and identifying actionable opportunities for cost saving in an effective way. In this paper, we proposed a comprehensive analytic pipeline for healthcare operational data and designed the Cost and Care Insight, an interactive and scalable hierarchical learning system for identifying cost saving opportunities, which provides attributable and actionable insights in improving healthcare management. This interactive system was built and tested on operational data from more than 750 facilities in ActionOI, a service assisting operational and performance evaluation in a realistic context. Here we introduce the design and framework of the system and demonstrate its use through a case study in the nursing department.

Keywords: Healthcare data mining · Cost saving opportunities · Value based care

1 Introduction

In recent years, more and more intelligent technologies and models have been performing an important role in healthcare. Much of this study is dedicated to helping patients [1, 2], some is disease-focused [3, 4], and some is geared toward healthcare organizations [5]. The techniques to help healthcare organizations are diverse, with studies targeting the management of healthcare costs becoming increasingly sophisticated. The value-based care paradigm necessitates hospitals and health systems to deliver on quality outcomes while improving operating margins which triggers healthcare providers to search for cost saving opportunities. However, this process requires massive manual efforts as the healthcare operation is a complex context and relies heavily on domain knowledge. In

most settings, hospitals tend to have separate management systems for different types of input—clinical, operational, and financial. The traditional approach to identify cost saving opportunities requires experts to integrate knowledge from different scales and drilling into the huge volume of data.

In recent studies on healthcare operation management, researchers have been trying to address this problem from policy level, individual level, hospital level and population level [6]. Aswani et al. [7] addressed this problem from policy level studying the Medicare Shared Saving Program (MSSP) and applied principal-agent model in optimizing Medicare saving. The Medicare Shared Saving Program (MSSP) was aiming to correct the misalignment of incentives between Medicare and providers by making it financially viable for providers to improve the efficiency of healthcare delivery. Providers who reduced their cost below the financial benchmark could receive bonus payment from Medicare. The proposed alternate approach could increase up to 40% Medical savings and took into account both the benchmark and the investment amount. As for individual level cost management, the emerging of EHR has accelerated the application of analytic methods in this area including applying k-nearest neighbors (KNN) in recommending personalized treatment for diabetes patients [8] and utilizing Lasso regression to effectively learn the optimal initial warfarin dose from high-dimensional covariates [9]. At hospital level, researchers have been studying the optimization problem focusing on staffing and resource scheduling in a specific setting like surgical department. Rath et al. proposed a two-stage robust mixed-integer system in improving operating cost and total cost by estimating the effect of surgical staffing decisions [10, 11]. Besides, Ang et al. addressed the hospital level operation management problem in a different way [12]. They focused on predicting emergency department (ED) wait time leveraging LASSO regression and a generalized fluid model of the queue, and the proposed model could reduce up to 30% mean squared error. At population level, Ta-Hsin et al. introduced a statistical process control (SPC) based detection algorithm to identify emerging healthcare cost drivers in a target population [13]. The hierarchical drill-down approach enables payers to early detect cost drivers within the same episode-based group.

Despite the promising value in identifying cost saving opportunities to reduce healthcare cost from different aspects, there are still many challenges to be solved. First, clinical and operational data is often separated in different systems. For instance, patients' data including individual conditions and outcomes is recorded in EHR system and hospital operational information like staffing and resource management structure is recorded in an isolated financial system, which made data integration and aggregation challenging. Second, the healthcare operational data is always high-dimensional and sparse which often contains as many as thousands [14] or tens of thousands [15] covariates. How to systematically process the noisy data and identify cost relevant features remains unaddressed. Third, the heterogeneity between and within healthcare organizations makes it impossible to develop a one-fit-all solution, while modeling for each scenario requires domain knowledge. Last but not the least, it is also difficult to identify root causes from a huge amount of variables, as well as to make sure the identified cost saving opportunities are actionable.

In this paper, we introduce a scalable hierarchical learning system for identifying cost saving opportunities, named ‘Cost and Care Insight’, which systematically addresses the

data analyzing pipeline and enables an interactive query. The system is built based on financial and operational data from over 750 facilities in US from ActionOI [16] system and provides cost saving opportunities from hospital level. Comparing to the traditional approach of making comparison with benchmarks, our system is able to automatically identify actionable features from high-dimensional data and analyze the attributable effect of each element towards cost leveraging statistical models such as the generalized estimate equation (GEE) model [17] and other state-of-art machine learning models.

2 Materials and Method

2.1 Data Source

The data utilized to develop the system was obtained from IBM ActionOI®, a service assisting operational and performance evaluation in a realistic context. It contains more than 750 healthcare organizations across the US. The operational and performance measures in ActionOI can be divided into 4 groups: (1) primary operational measures provided by hospital such as organization chart, staffing, scheduling and device-maintained information; (2) financial measures such as operating costs, operating beds and patient discharges; (3) secondary measures calculated from the primary data, e.g. labor expense per patient day and operating expense per 100 devices maintained; (4) performance measures including patient average length of stay etc. (5) fixed operational characteristics, e.g. hospital major teaching status, workload adjustment factors, area wage index (AWI), level 1 trauma center etc.

The measures are heterogeneous among hospitals or even departments within one facility, leading to over 12,000 different features recorded in the system. A snapshot of the most frequent reported measures in nursing service is displayed in Table 1.

Table 1. Typical reported measures in nursing service in our collected data.

Feature type	Example	Data type
Primary operational measures	Worked Hours: Staff	numeric
Financial measures	Labor Expense non-MD per patient day	numeric
Secondary measures	Skill Mix: RN%	numeric
Performance measures	Patient average length of stay	numeric
Fixed characteristics	Hospital major teaching status	character

The primary data are expected to be quarterly reported to ActionOI system, while in reality, it tends to have many missing values and noisy records in the first 3 quarters.

2.2 Overall Architecture

A traditional and straightforward approach in analyzing healthcare performance metrics is to compare with benchmarks. ActionOI provides the comparison between each hospital with top hundred facilities for each element of interest. However, this labor-intensive

approach lacks a systematic overview and data scientists can be overwhelmed by the sheer volume of elements.

Here we introduce Cost and Care Insight for automatically identifying cost saving opportunities. The system contains an offline training engine and an online query engine as shown in Fig. 1.

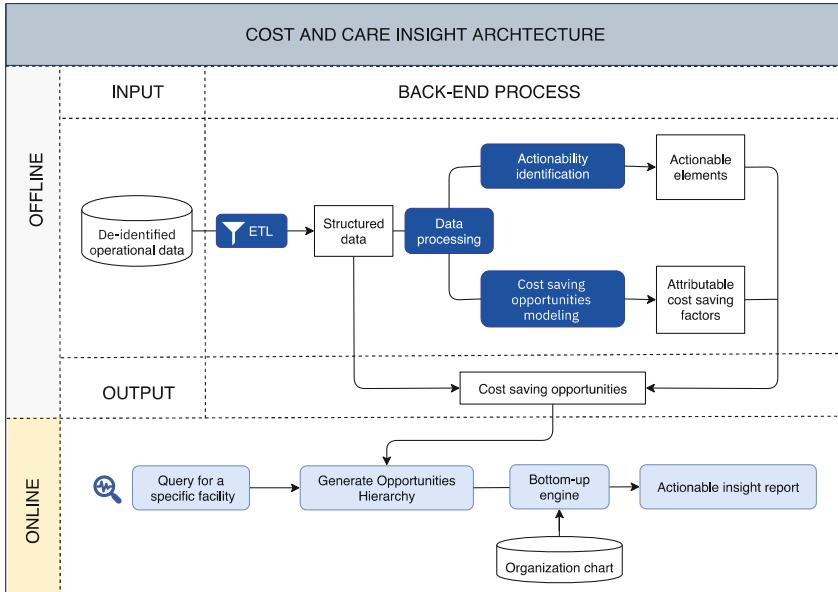


Fig. 1. Overall architecture of Cost and Care Insight.

The offline engine takes the de-identified operational data from financial system and EHR system as input. Then the engine will generate cost saving opportunities from the most granular level through the back-end process which contains four major modules for coordinating the tasks, including ETL module, Data processing module, Actionability identification module and a Cost saving opportunity modeling module. The goal of the online query engine is to enable the user to visualize the cost structure and opportunities for a targeting facility. The users first input a query, then the result of targeting facility will be extracted from the offline output, to form an opportunities hierarchy. The next step is a bottom-up engine which leverages the hierarchy and the organization chart of the specific facility and finally generate an actionable insight report in an interactive approach.

The detailed description of the major components can be found in the following sections.

ETL Module. The goal of this module is to accelerate data extraction from the operational database. To implement this, we designed a spark compatible approach using pyspark in python. The ETL tools first summarize the facility number for each target,

then extract elements with at least 300 facilities ever reporting it. There could also be multiple optimizing targets in each service line including labor expense, supply expense and total expense etc., and we take 500 facilities as a threshold for deciding sub tasks(regard as a sub task if more than 500 facilities report the target in a specific service line). Finally, 86 sub tasks were identified using this criterion. Table 2 listed 12 tasks that are most frequently reported.

Table 2. Summary of optimizing targets in 12 most frequently reported service line.

Service line	Optimizing target	Number of facilities	Number of features
Nursing	Labor Expense per Equivalent Patient Day	887	192
Imaging	Labor Expense per APC Relative Weight	832	190
Laboratory	Labor Expense per 100 Billed Tests	771	284
Surgical	Expense \$ per operating room case: Labor	886	172
Respiratory and Pulmonary Care	Labor Expense per APC Relative Weight	796	142
Emergency Facility	Labor Expense per APC Relative Weight	761	158
	Labor Expense per 1000 Gross Square Feet Maintained	850	123
Supply Chain	Labor Expense per SIS Weighted Adjusted Discharges	569	157
Pharmacy	Labor Expense per CMI Weighted Dept Adjusted Discharge	805	194
Other Support	Labor Expense per 1000 Gross Square Feet Patrolled	691	114
Revenue Cycle Management	Labor Expense per 100 Patient Registrations	654	103
Clinical Resource Management	Labor Expense per Admission and Registration	525	130

* APC: Ambulatory Payment Classifications. The APC is the service classification system for the outpatient prospective payment system, and the APC relative weight measures the resource requirements of the service and is based on the geometric mean cost of services in that APC. SIS: Supply Intensity Score. CMI: Case Max Index, represents the average diagnosis-related group (DRG) relative weight for that hospital.

Take nursing service as an example, nursing departments, we are more interested in saving labor expense, and the most frequently reported target is *labor expense per*

equivalent patient day with 887 facilities having been reporting it. Then ETL performed feature extraction from nursing departments and found total 192 features recorded in at least 300 facilities.

The output of ETL module are 86 datasets for 12 service lines which are aggregated yearly (taking the Q4 data as annual report).

Data Processing Module. The operational data is very sparse and noisy. Hence, we introduce 3 sub-modules including a) extreme value detection, b) imputation and c) feature selection.

Extreme Value Detection: One third of the features included in our study are categorical (binary) and the rest are continuous variables. The process of extreme value detection for continuous variable is illustrated in Fig. 2A. The first step is to decide whether there is a prior threshold for the given variable. For example, variables that are related to staff proportion and cost proportion, such as skill mix: RN% have prior threshold [0, 100] Fig. 3A. The process removes the extreme value beyond the threshold and goes to the step 2, detecting massive zeros in the distribution. We observe a massive number of zero values in many variables and find out it might be human error when tracing back (see example in Fig. 3B). So, we designed this density-based approach to solve this problem where we calculate the difference between the original density curve of x' with the density curve after median filtering. The massive zeros are identified if a spike is discovered in its distribution:

$$\maxima(\text{curve}_a - \text{curve}_b) > \theta \quad (1)$$

and

$$\text{index}(\maxima((\text{curve}_a - \text{curve}_b))) = 0 \quad (2)$$

where

$$\theta = 10 * \text{length}(x) / \text{count}(\text{unique}(x)) \quad (3)$$

For variables without massive zeros, we apply Log transform otherwise the Inverse hyperbolic sine (IHS) transform [18] is applied to reduce the skewness:

$$\text{arcarsinh}(x) = \log(x + \sqrt{x^2 + 1}) \quad (4)$$

Then extreme values outside Tukey fences (3IQR) are removed (the Tukey's method for outlier detection).

Data Imputation: After removing the extreme values, the processing module will continue to conduct data imputation. Several general methods of imputation are deployed in our system including mode, median, zero, forward, k nearest neighbor (KNN) and Multiple Imputation (MI). The default method for binary variables and continuous variables are mode imputation and median imputation respectively.

Feature Selection: We implement a feature selection module for confounder reduction with several feature selection methods in the data processing library including step-wise regression, tree-based method and lasso regression.

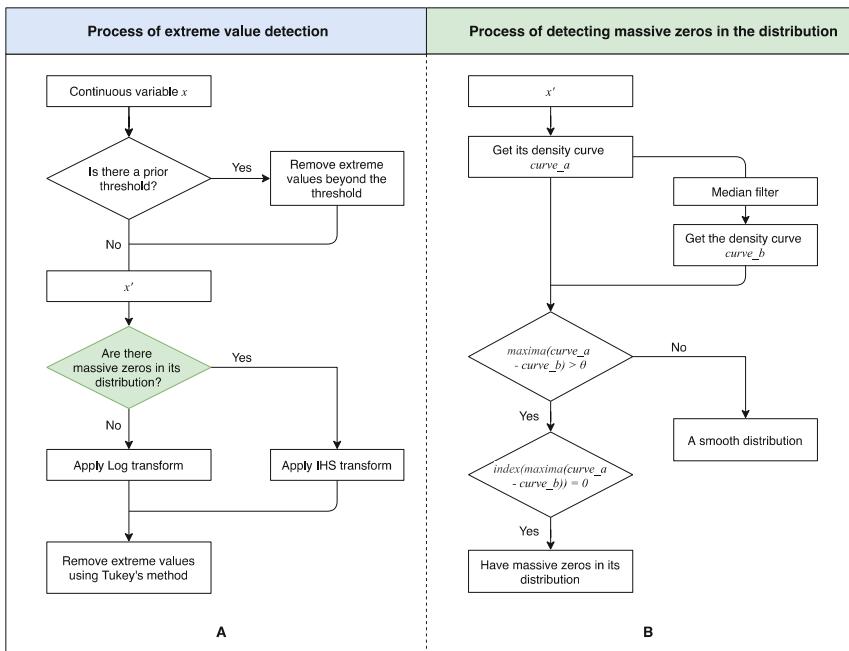


Fig. 2. Extreme value detection for continuous variables. A). the process of extreme value detection. B). the process of detecting massive zeros in the distribution

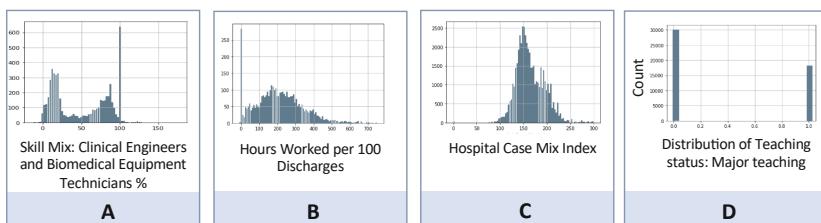


Fig. 3. Distribution of four types of measures. A). a feature with a fixed threshold. B).a feature with massive zeros. C). a feature of log-normal distribution. D). a feature of Bernoulli distribution

Actionability Identification Module. The key idea to identify the actionable factors is based on the historical trend. We decomposed each feature to the most fine-grained elements and determine the actionability for each element. If any of the components is actionable then the feature is regarded as actionable. The actionability is defined as:

$$\alpha_i = \begin{cases} 0, & \text{if } \sum_1^n \{1 | \sum_1^t (y_{it} - \bar{y}_i)^2 = 0\}_i \geq 0.9 * n \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

The actionability can be different across facilities, so the system is designed also providing a manual adjustment feature actionability.

Cost Saving Opportunity Modeling Module

The Baseline Model: GEE In the spirit of a common recommendation on prediction for longitudinal data analysis, the baseline model start with the generalized estimation equation (GEE) model, which is a general statistical model to fit a marginal model for longitudinal data. It provides the population-average estimates in response over time adjusting for covariates. Given a longitudinal/clustered data set for K individual, suppose there are n_i observations for each individual i , $i = 1, 2, \dots, K$, let Y_{ij} , $i = 1, 2, \dots, n_i$ denotes the j^{th} response from the i^{th} subject and $X_{ijp} = x_{ij1}, x_{ij2}, \dots, x_{ijp}$ denotes the vector of p covariates. The marginal expectation of $E(Y_i|X_i) = \mu_i$ is modelled by:

$$g(\mu_{ij}) = X'_{ij\beta} \quad (6)$$

where g is known link function, and $\beta^T = (\beta_1, \dots, \beta_p)^T$ is a $p - dimensional$ vector of regression coefficient needs to be estimate.

GEE uses a variance-covariance matrix to measure the within subject variance of repeated measurements for the same subject. Using the standard notation is proposed by Liang and Zegar [19], the working covariance matrix is denoted by $V_i = A_i^{1/2} M_i(\alpha) A_i^{1/2}$, where $A_i = Diagv(\mu_i, \dots, v(\mu_{ij}))$ is the diagonal matrix with the known variance function $v(\mu_{ij})$ and $M_i(\alpha)$ is the corresponding working correlation matrix, and α is unknown. The GEE estimated the regression coefficient by the following equation:

$$U(\beta, \alpha) = \sum_{i=1}^N D_i^T V_i^{-1} (Y_i - \mu_i) = 0 \quad (7)$$

where $D_i = \partial \mu_i / \partial \beta'$. By using a sandwich estimator, GEE yields a robust estimation even when the correlation structure is mis-specified [19].

Model Evaluation: The default baseline model is GEE with auto-regressive covariance structure which assume the correlation between time points within each facility decreases as a power of how many timepoints apart two observations. We evaluate the model performance by splitting the dataset in 3:1 ratio with regarding to the reporting year. The root mean square error (RMSE), mean absolute error (MAE) and R^2 are calculated for training and testing set and we compare the performance of choosing for different working covariance structure in GEE model including independence, exchangeable and stationary working covariance structures. Besides, other advance regression models including support vector regression, random forest regression (PCR), deep neural network regression and LightGBM regression are also adopted in our model library and are compared with the baseline model.

Bottom-Up Strategy: We attribute the cost impact of each operational metric by multiply the difference of actual value and expected value of each metric/measurement by correspondent coefficient. The cost saving opportunities in the most granular level are calculated as:

$$g(x_{ij}) = \alpha_{ij} * \omega_i(x_{ij} - E(x_{ij})) \quad (8)$$

where x_{ij} represents the i^{th} measure of the j^{th} department, α_{ij} is the actionability of each element, ω_i is the weight calculated from the regression model, and $E(x_{ij})$ is the expected value of each operational metrics/measurement i within the same department j .

For each level in the organization hierarchy, total cost saving opportunities at that level is calculated in a bottom-up manner by summing up all actionable cost impact with positive values:

$$\sum g(x_{ij})|g(x_{ij}) > 0 \quad (9)$$

3 Results

In this section, we will describe a case study for nursing services in ActionOI to demonstrate the use of our system. We evaluate the model performance comparing four different working correlation structures in GEE model as well as with other regression methods including support vector regression, random forest regression, deep neural network regression and LightGBM regression.

From 2015 to 2018, there are totally 784 facilities providing nursing service in 51 unique departments collected in ActionOI. Targeting at identifying cost saving opportunities for labor expense per equivalent patient day in nursing services, we extracted 192 most frequently reported features from 12080 measures (only those measures that appear in at least 300 facilities will be extracted). In the baseline GEE model, 69 features are selected by step-wise regression and 10 of them are identified as actionable cost saving factors as shown in Table 3. The following elements are significantly associated with lower labor cost with p value under 0.05, including hire new graduates from on-site internship program, charge nurse provides clinical patient care 50% or more of the time, having 80% or more full time status employees, skill mix percentage of RN and hours worked as % of hours paid. Besides, those elements including overtime hours as % of worked hours of non-MD, hours paid per equivalent patient day, skill mix percentage of other patient care and contract hours paid as % of total hours paid have significant positive correlation with labor expense. These elements are regarded as having potential cost saving impacts and are then put into in the cost opportunity equation to calculate the final opportunity volume.

Table 3. Actionable features identified through GEE model.

Feature name	Coefficient	P value
Hire new graduates from an on site internship program?	-11.07	<0.001
Charge nurse provide clinical patient care > 50%?	-8.82	0.025
Have 80% or more full time status employees?	-17.75	<0.001
Overtime Hours as% of Worked Hours: Non MD	4.19	<0.001
Hours Paid per Equivalent Patient Day	21.89	<0.001
Skill Mix: Other Patient Care %	0.60	0.042

(continued)

Table 3. (continued)

Feature name	Coefficient	P value
Skill Mix: RN %	-2.25	0.016
Contract Hours Paid as % of Total Hours Paid	2.71	<0.001
Worked Hours: Non MD	-0.001	0.009
Hours Worked as % of Hours Paid	-6.99	<0.001

* FTE: full time equivalents

Combining the actual value of each facility and their organization chart, the system can generate the cost structure and cost saving opportunities at each organization level as shown in Fig. 4. The interface enables users to choose the facility, department and query period of their interest and specify a target to optimize. For example, in the cost structure for facility 759, three elements are identified as cost saving opportunities: skill mix percentage of management, hours paid per equivalent patient day and hours worked as % of hours paid which have opportunity for saving \$11,705.93, \$480,097.89 and \$49,076.88 respectively. The bottom-up strategy is adopted which aggregates the opportunities from the most granular level to upper levels. A total of \$3.73 million saving is identified for nursing services summing up all care units and a total of \$7.18 million for all service lines is identified in facility 759.

The visualization tab provides advance comparisons within and between facilities. As shown in Fig. 5, the opportunities across all facilities in the chosen period is displayed where the actual labor expense is also shown for reference. The three plots below are yearly trend of total cost saving opportunity in the chosen facility, opportunity for each department within the facility and yearly opportunity trend of a chosen element of interest. These plots together depict the actionable cost saving opportunity insight from the most granular level to upper levels for the specific facility of interest.

Table 4 shows the evaluation performance of each model. The default model adopted in this system is GEE model with auto-regressive working correlation structure which makes an assumption that the time series data within each facility are correlated and the correlation decreases as a power of how many timepoints apart two observations are. It has achieved a RMSE of 170.28, MAE of 89.53, R^2 of 0.81 in the training set and RMSE of 188.89, MAE of 96.86, R^2 of 0.79 in the testing set. The GEE model with independence covariance structure equals the linear regression which assumes the time series data within each facility is independent. It achieves a lower R^2 of 0.79 in testing set but higher RMSE and MAE of 188.89 and 96.86 comparing to the baseline model. The exchangeable covariance structure assumes the correlation stays the same over time and the stationary covariance structure assumes a stationary time series which the correlation is only related to the time interval but not the time point. These two models achieved a better MAE and R^2 but higher RMSE in the testing set.

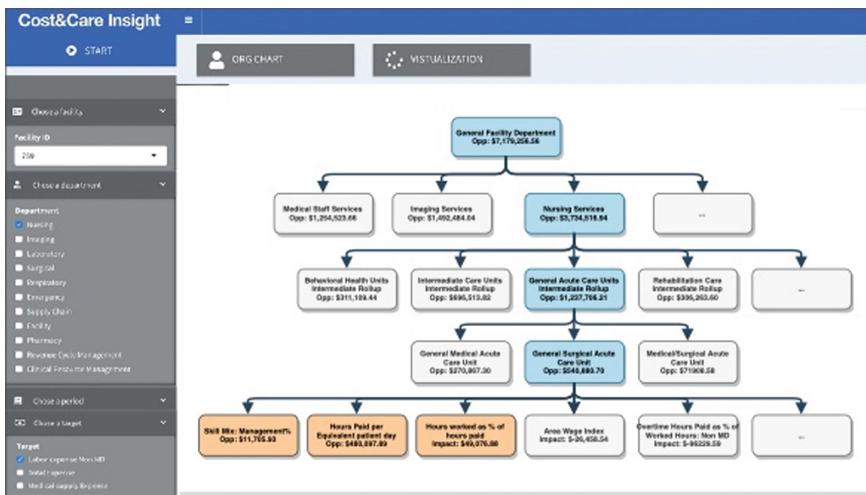


Fig. 4. User interface of Cost and Care Insight system.

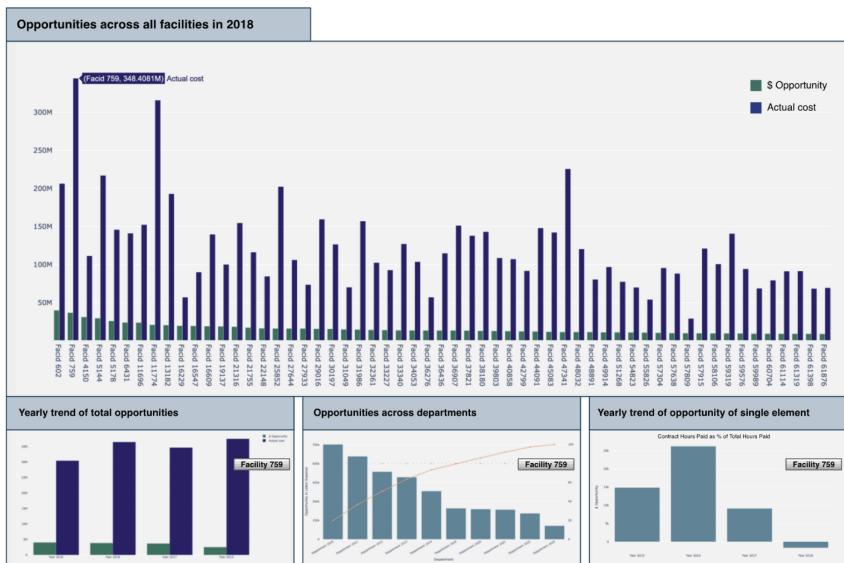


Fig. 5. Advance comparison in visualization tab.

Advance machine learning models are also included in our model library, and the result shows LightGBM outperforms all other models in testing data set with RMSE of 138.47 and R^2 of 0.89. Random forest regression achieved the lowest MAE and Deep regression achieved a second higher R^2 of 0.88.

Table 4. Model evaluation result.

Model	RMSE		MAE		R^2	
	Training	Testing	Training	Testing	Training	Testing
GEE(ACS)	170.28	188.89	89.53	96.86	0.81	0.79
GEE(ICS)	166.88	185.66	89.45	95.93	0.82	0.80
GEE(ECS)	174.95	193.96	87.16	92.21	0.80	0.78
GEE(SCS)	167.32	186.05	88.80	95.54	0.82	0.80
Support Vector Regression	385.81	421.45	223.91	238.72	0.02	0.03
Random forest regression	49.35	153.54	20.22	62.75	0.98	0.86
Deep regression	119.76	146.69	50.30	65.24	0.91	0.88
LightGBM	86.01	138.47	44.17	63.33	0.95	0.89

* RMSE: Root Mean Squared Error * MAE: Mean Absolute Error * R^2 : Coefficient of determination * ACS: Autoregressive covariance structure * ICS: Independence covariance structure * ECS: Exchangeable covariance structure * SCS: stationary covariance structure

4 Discussion

Healthcare cost management has drawn much attention in the time of value-based care. There are many challenges in analyzing healthcare operational performance and it has not yet been fully addressed the general pipeline of identifying cost saving opportunity. To address the unmet needs in healthcare performance analysis from hospital level, we proposed the Cost and Care Insight system, which is designed based on real world needs and scenario.

The general framework enables the cost saving opportunity discover in a hierarchical manner, however, there are also some unaddressed needs and worth further implement. First, the system utilizes statistical and machine learning approaches but did not include the causal inference to locate the root causes. Second, we provide a general method to identify the actionable features based on the historical trend, however, the actionability of each element differs across facilities. For example, for hospitals don't support a level I trauma center, some would think it's too costly and not actionable in the short term, but some would regard it as actionable if it helps improving healthcare quality and actually saves money in the long run. So, for now we leave a choice for users to make adjustment in the current system, but in the future, a further implement is needed to provide both short term advice and long-term advice.

We are aware that this system is far from optimal and can be improved from many perspectives. The future work will focus on introducing causal inference in our library to provide what-if analysis and implementing actionability identification in a more flexible way.

5 Conclusion

In this paper, we proposed an interactive and scalable hierarchical learning system for identifying actionable cost saving opportunities and systematically addressed the analytic pipeline for healthcare operational performance analyses. The hierarchical learning system provides: 1. Actionable insights identified from massive operational measurements. 2. Attributable cost saving opportunities identified by proper statistical models and state-of-art machine learning approaches to supplement and consolidate the findings. 3. A spark compatible architecture and a customized interactive interface which enables more flexible query and display.

To the best of our knowledge, this is the first time the complete process of analyzing healthcare operational data from hospital level has been fully addressed. The system enables actionable insight discover and has the potential to improve healthcare performance management in the time of value-based care.

Acknowledgement. The author wise to thank D.K for his expertise on dealing with healthcare operational data and his advice on study design as well as data analyses. The author also thanks M.J for her guidance in this project.

References

1. Lin, C., et al.: SenseMood: depression detection on social media. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 407–411 (2020)
2. Wang, Y., et al. Automatic depression detection via facial expressions using multiple instance learning. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1933–1936. IEEE (2020)
3. Tang, Z., et al.: Embracing disease progression with a learning system for real world evidence discovery. In: Huang, D.-S., Jo, K.-H. (eds.) ICIC 2020. LNCS, vol. 12464, pp. 524–534. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60802-6_46
4. Hu, P., et al.: Predicting hospital readmission of diabetics using deep forest. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI), pp. 1–2. IEEE (2019)
5. Fainman, E.Z., Kucukyazici, B.: Design of financial incentives and payment schemes in healthcare systems: a review. Socio-Econ. Plan. Sci. **72**, 100901 (2020). <https://doi.org/10.1016/j.seps.2020.100901>
6. Mišić, V.V., Perakis, G.: Data analytics in operations management: a review. Manuf. Serv. Oper. Manag. **22**(1), 158–169 (2020)
7. Aswani, A., Shen, Z.J.M., Siddiq, A.: Data-driven incentive design in the medicare shared savings program. Oper. Res. **67**(4), 1002–1026 (2019)
8. Bertsimas, D., Kallus, N., Weinstein, A.M., et al.: Personalized diabetes management using electronic medical records. Diabetes Care **40**(2), 210–217 (2017)
9. Bastani, H., Bayati, M.: Online decision making with high-dimensional covariates. Oper. Res. **68**(1), 276–294 (2020)
10. Rath, S., Rajaram, K., Mahajan, A.: Integrated anesthesiologist and room scheduling for surgeries: methodology and application. Oper. Res. **65**(6), 1460–1478 (2017)
11. Rath, S., Rajaram, K.: Staff planning for hospitals with cost estimation and optimization. Kenan Institute of Private Enterprise Research Paper, pp. 18–28 (2018)

12. Ang, E., Kwasnick, S., Bayati, M., et al.: Accurate emergency department wait time prediction. *Manuf. Serv. Oper. Manag.* **18**(1), 141–156 (2016)
13. Li, T.H., Jiang, H., Tran, K., et al.: A Systematic Approach to Detect Hierarchical Healthcare Cost Drivers and Interpretable Change Patterns. arXiv preprint [arXiv:1907.08237](https://arxiv.org/abs/1907.08237) (2019)
14. Bayati, M., Braverman, M., Gillam, M., et al.: Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS One* **9**(10) (2014)
15. Razavian, N., Blecker, S., Schmidt, A.M., et al.: Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* **3**(4), 277–287 (2015)
16. Wang, B., Eliason, R.W., Richards, S.M., et al.: Clinical engineering benchmarking: an analysis of American acute care hospitals. *J. Clin. Eng.* **33**(1), 24–27 (2008)
17. Feng, Z., Diehr, P., Peterson, A., McLerran, D.: Selected statistical issues in group randomized trials. *Annu. Rev. Public Health* **22**, 167–187 (2001)
18. Bellemare, M.F., Wichman, C.J.: Elasticities and the inverse hyperbolic sine transformation. *Oxford Bull. Econ. Stat.* **82**(1), 50–61 (2019). <https://doi.org/10.1111/obes.12325>
19. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986)



A Sub-network Aggregation Neural Network for Non-invasive Blood Pressure Prediction

Xinghui Zhang¹, Chunhou Zheng^{1(✉)}, Peng Chen^{2(✉)}, Jun Zhang³, and Bing Wang⁴

¹ School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China

² Institutes of Physical Science and Information Technology and School of Internet, Anhui University, Hefei 230601, Anhui, China
pengchen@ustc.edu

³ School of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, China

⁴ School of Electrical and Information Engineering, Anhui University of Technology, Hefei 243032, Anhui, China

Abstract. Non-invasive blood pressure prediction is an important method to prevent diseases such as hypertension. This paper proposes a sub-network aggregation with large convolution kernel convolution to predict non-invasive blood pressure. First, the large convolution kernel module in the backbone network is used to extract PPG data features. Then, the multi-scale features of the backbone network are fused and then aggregated with the features extracted from the parallel sub-network. Finally, ABP data is predicted by convolution, and then blood pressure is predicted according to the relationship between ABP data and blood pressure. In this work, the large convolution kernel is used to extract more information, and the feature extraction of subnetwork is used to help the prediction of backbone network, which further achieves improved prediction. Under the BHS standard, the prediction accuracy of blood pressure based on DBP and MAP can reach grade A. In addition, the prediction accuracy of DBP and MAP can also reach the standard in terms of AAMI standard.

Keywords: Subnetwork aggregation · Large convolution kernel · Backbone network · PPG data · ABP data

1 Introduction

Research [1] shows that global cardiovascular disease and mortality have increased since 1990. In 2019, the number of people suffering from cardiovascular diseases has reached 523 million, and the number of deaths has reached 18.6 million. Hypertension is the most important cause of death in cardiovascular diseases [2]. Non-invasive blood pressure prediction has become one of the most important tasks in the prevention of cardiovascular diseases.

In the field of non-invasive blood pressure prediction, the main research can be divided into physiological models and regression models. The physiological model mainly uses the physiological characteristics PTT (Pulse transit time) and PWV (Pulse

wave velocity) for prediction. PTT or PWV features are generally extracted from ECG (Electrocardiogram) and PPG (Photoplethysmograph) or PPG and PPG at the same time. The related research is as follows. Chan et al. [3] extracted PTT features from PPG and ECG data, and used a linear model to predict blood pressure. Poon et al. [4] proposed a non-linear model and obtained the prediction result that the difference between SBP (Systolic blood pressure) and blood pressure is 0.6 ± 9.8 mmHg, which has the potential to be applied to wearable devices. Yan et al. [5] found that PTT data and DBP (Diastolic blood pressure) have a strong correlation, which leads to a theoretical foundation for the non-invasive blood pressure prediction method based on PWV.

The regression method for blood pressure prediction is to take ECG or PPG data as input. Monika et al. [6] only took ECG data as input and applied machine learning methods to predict the regression value of blood pressure, which achieved a score of 13.52 mmHg for MAE (Mean Absolute Error) of MAP (Mean arterial pressure). Mousavi et al. [7] extracted appropriate feature vectors in the frequency domain and used machine learning methods to predict blood pressure. Suzuki and Oguri [8] proposed a random forest algorithm to predict blood pressure, and obtained a result with an average error of 1.2 mmHg. Later, El-Hajj and Kyriacou [9] pointed out that random forest is not suitable for handling complex tasks. Research work [10–12] used feedforward neural network, SVM combined with discrete wavelet transform, and least square regression for non-invasive blood pressure prediction. Brophy et al. [13] uses Unet [14] model to estimate ABP (Arterial blood pressure) data based on PPG data to predict blood pressure. Because the relationship between ABP data and blood pressure can be expressed as:

$$\text{MAP} = \text{mean}(\text{ABP}), \text{SBP} = \max(\text{ABP}), \text{DBP} = \min(\text{ABP}). \quad (1)$$

In order to ensure the extraction of reliable physiological features, the method of physiological model requires manual verification every period of time. Compared with the physiological model, the regression model has lower data requirements and does not require manual verification. The ECG data collection operation in the regression model is complicated, while the PPG data collection is relatively simple, and it only needs the collector to be close to the wrist. Taking into account the need to continuously monitor blood pressure and the complexity of future scenarios, this paper proposes a regression model method based on PPG data only to predict blood pressure.

2 Method

2.1 Model

The overall structure of the proposed model in this paper is shown in Fig. 1. It includes two Unet structures, backbone network Unet4 for MaxPooling1D 4 times and subnet Unet1 for MaxPooling1D once. First, the backbone network Unet4 learns data features for quick prediction. Then, to improve the prediction accuracy, the sub-network Unet1 is used to extract different data features. Finally, the extracted features of the two networks are aggregated and passed through a one-dimensional volume. The product process gets the final prediction result.

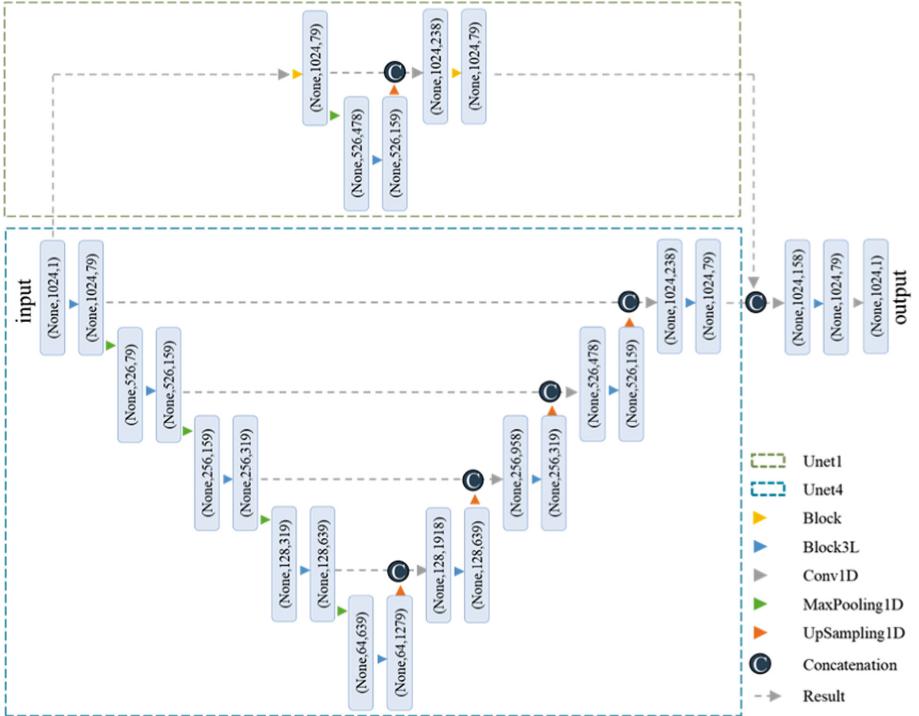


Fig. 1. Overall structure of model

2.2 Backbone Network

The backbone network Unet4 of the model is shown in Fig. 1. The network structure is composed of coding structure, layer jump connection structure and decoding structure. The maxpooling part on the left is the coding structure, which reduces the length of data through convolution and maxpooling, and extracts shallow features. The up-sampling part on the right is the decoding structure, which obtains in-depth features through convolution and upsampling. The middle jump layer connection structure concatenates the results obtained from the encoding stage and the decoding stage through Concatenate function. In this way, the output result not only has deep-level features, but also retains shallow-level feature information.

2.3 Large Convolution Kernel

The main module Block3L of the model is shown in Fig. 2. It is the main module of the Unet4 structure. The Block3L module first convolves the input data of the module with the kernel size of 3, 5, and 7 in order to extract features; then uses the kernel size of 1 to perform one-dimensional convolution on the input data of the module; finally, the results of the two parts are added to get the final result.

Block is the main module of Unet1 structure, which is similar to Block3L. The only difference is in that the kernels of Block are all 3, while the kernels of Block3L

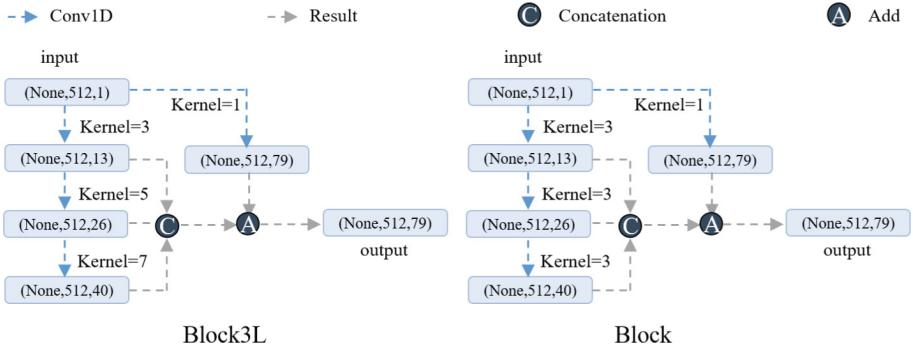


Fig. 2. The main modules of the model are Block3L and Block

are 3, 5, and 7, respectively. The Block3L module has 3 obvious characteristics, which are also the main reasons why the Unet4 structure adopts Block3L module. The three characteristics are:

- 1) The convolution kernels are all odd numbers. For images, the odd number of kernels makes filling easier. Assuming that the size of the image is $n * n$, the size of the convolution kernel is $k * k$, and the padding amplitude is $(k - 1)/2$. The output obtained after convolution is $(n - k + 2 * (k - 1)/2)/1 + 1 = n$, that is, the output image size is also $n * n$.
- 2) The convolution kernels are all different in size. Szegedy et al. [15] pointed out that two $3 * 3$ convolution kernels and a $5 * 5$ convolution kernel have the same receptive field, that is, the larger convolution kernel has a larger receptive field. Different receptive fields yield different results, and usually the results obtained by large receptive fields can contain more information.
- 3) The convolution kernels are all small in size. Convolution kernel greater than 1 can improve receptive field, so usually the convolution kernel for feature extraction will not be set to 1. Moreover, the use of large-size convolution kernels will relatively extend the calculation time, so small-size convolution kernels are often used.

2.4 Sub-network Aggregation

The model proposed in this paper is based on sub-network aggregation to improve the prediction accuracy, and the parallel connection of the sub-network and the backbone network to speed up blood pressure prediction. Sub-network aggregation mainly uses the high-level features extracted from the backbone network to aggregate the low-level features of the sub-network to improve the expression ability of mixed features and refine prediction results.

This paper designs the sub-network aggregation structure to ensure the output of the backbone network. At the same time, this paper did not choose GoogleNet [16] and other large-scale network models as the model backbone, because large-scale models are computationally expensive, for example GoogleNet's calculation amount is 1500

MFLOPs [17]. Therefore, this work does not choose the method of large-scale model, which can relatively reduce the calculation time and achieve the effect of rapid prediction.

3 Results

3.1 Dataset

The data used in the experiment is a simple pre-processed data set collected from MIMIC II database [18] in literature [19]. This data set mainly contains PPG, ABP and ECG data. All three types of data are sampled at a frequency of 125 Hz, and only PPG and ABP data are used in this work. The true value SBP, DBP and MAP in the data set are the maximum, minimum and average values of each set of sampling points of the true value ABP.

The selected SBP range is [71 mmHg, 199 mmHg], and the variance of SBP is 527.74 mmHg. The selected DBP range is [50 mmHg, 174 mmHg], where the variance of DBP is 114.01 mmHg. It can be seen that the variance of SBP is much larger than that of DBP.

3.2 Experimental Details

The dataset is composed of 127260 groups of data, where each group has 1024 one-dimensional data. First of all, this work took the PPG data with simple noise removal as input, and trained on the simple Unet network with the MAE loss function to obtain the rough accuracy data of ABP. Then, the ABP data with rough accuracy was used as the input of the proposed model, and the final ABP prediction model was trained with the MSE (Mean Square Error) loss function. Finally, the corresponding SBP, DBP and MAP data were calculated based on the predicted ABP data. The training method was used in this work because the following evaluation methods need to use ME (Mean Error) and MSE.

Here 100,000 samples were selected randomly for training the proposed model and the remaining data were used as the test set, where 90,000 samples were used as the training set and 10,000 samples were used as the validation set. The Training on the proposed model is performed in a 10-fold cross-validation technique, and Adam algorithm was used as optimization function.

3.3 Compare with Existing Methods

3.3.1 BHS Standard

The BHS standard is a scheme proposed by the British Hypertension Society to evaluate the accuracy of blood pressure measurement method. There are grades A, B and C in the BHS standard. One method will have grade A if it predicts that the blood pressure error value within 5 mmHg accounts for more than 60% predictions, while the error value within 10 mmHg accounts for more than 85% predictions, and the error value within 15 mmHg accounts for more than 95% predictions. Grade B and C can be calculated

Table 1. Comparison of different methods in terms of prediction accuracy and cost time

		Accuracy			Time (ms)
		$\leq 5 \text{ mmHg}$	$\leq 10 \text{ mmHg}$	$\leq 15 \text{ mmHg}$	
BHS	Grade A	60%	85%	95%	/
	Grade B	50%	75%	90%	
	Grade C	40%	65%	85%	
LSTM	DBP	/	/	/	83.77
	MAP				
	SBP				
GRU	DBP	/	/	/	68.32
	MAP				
	SBP				
PPG2ABP	DBP	80.33%	92.03%	95.76%	2.66
	MAP	86.58%	94.68%	97.49%	
	SBP	68.91%	83.21%	89.62%	
MobileNet	DBP	77.49%	89.64%	94.18%	1.84
	MAP	86.65%	94.61%	97.62%	
	SBP	51.68%	74.13%	83.73%	
SqueezeNet	DBP	78.52%	89.50%	93.91%	2.83
	MAP	86.94%	94.67%	97.53%	
	SBP	56.82%	77.51%	86.24%	
Block1	DBP	82.46%	92.56%	95.81%	2.61
	MAP	87.30%	94.81%	97.53%	
	SBP	71.47%	85.50%	91.00%	
Block2	DBP	83.37%	92.99%	96.21%	2.19
	MAP	87.44%	94.89%	97.57%	
	SBP	72.21%	85.63%	91.08%	
Our Method	DBP	84.01%	93.11%	96.28%	2.67
	MAP	87.49%	94.90%	97.55%	
	SBP	72.98%	85.93%	91.23%	

from Table 1 in the same way. Table 1 lists the results of the BHS standard and existing method experiments, where “/” means unknown.

It can be seen from Table 1 that LSTM takes a long time to predict, and it can be seen that it is not suitable for applications with strict requirements on time. Even using the GRU network, the prediction time is 25 times slower than PPG2ABP. Therefore, the results of the LSTM and GRU algorithms is no longer compared.

MobileNet and SqueezeNet are lightweight neural networks, which run fast. It can be seen in Table 1 that although the lightweight neural network has short prediction time, it indeed achieves lower prediction accuracy than the algorithm in this work. At the same time, SqueezeNet has a 1.3% higher prediction accuracy than MobileNet. The difference between the two is that the former has an up-sampling process after maximum pooling. This also shows that Unet's pooled up-sampling scaling operation can indeed get more information.

In Table 1, Block1, Block2 and our method are respectively ablation experiments. The only difference between Block1 and PPG2ABP is that the former uses Block3L as the main body. The only difference between Block2 and Block1 is that the former takes the Unet4 structure as the main body. The only difference between our method and Block2 is that our method has a Unet1 subnet. It can be seen from the data that the integration of Block3L module, Unet4 network and Unet1 sub-network can indeed improve blood pressure prediction.

3.3.2 Bland-Altman Analysis

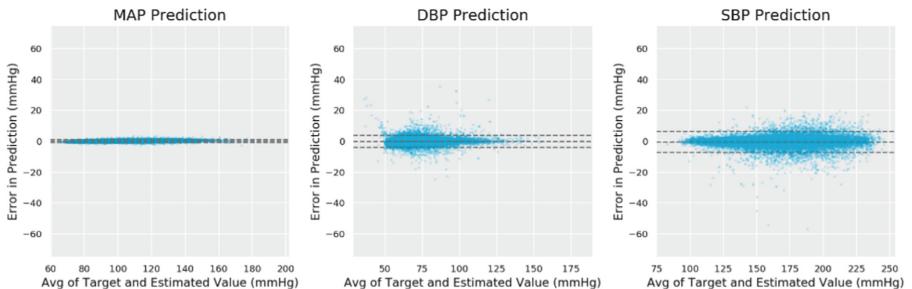


Fig. 3. Scatter plot analyzed by bland-altman

The picture drawn in Fig. 3 is a scatter plot commonly used in the Bland-Altman analysis method. The Bland-Altman analysis method [22] was proposed by Bland and Altman. It is also called the consistency evaluation of the continuity index. This method is a relatively new method that can replace the existing method. The new method here refers to the method proposed in this article, and the old method refers to the PPG2ABP method. The horizontal axis in Fig. 3 is the mean value of the difference between the prediction result and the real blood pressure value, and the vertical axis is the standard deviation of the difference between the prediction result and the real blood pressure value. At the same time, the 95% consistency limit is marked in the figure, and the upper and lower limits are $ME + 1.96 * STD$ and $ME - 1.96 * STD$, respectively.

When the percentage of the number of differences in the interval to the total number is not less than 95%, it can be considered that the new method can replace the old method. It can be seen from the figure that most of the values are within this interval. Moreover, in the experimental results, the percentages of SBP, DBP, and MAP in this interval are exactly equal to 95%. The results show that the method proposed in this paper can theoretically replace the method of PPG2ABP.

4 Conclusion

The experimental results of the proposed method in this paper can reflect the feasibility and superiority of large convolution kernel and sub-network aggregation in blood pressure prediction tasks. One advantage of the method is that the forecast time is short enough, which is at least 25 times faster than the LSTM or GRU algorithm. Another advantage is that compared with lightweight neural networks, the method performs better. The model in this paper undergoes multiple maximum pooling and up-sampling processes, which can fuse information with different sizes to improve the accuracy of blood pressure prediction. In the MIMIC II data set, this work only uses PPG data to predict blood pressure and obtain good results. Under the BHS standard, the blood pressure prediction of the method in terms of DBP and MAP reaches level A. In addition, under the AAMI standard, the blood pressure prediction of the method in terms of DBP and MAP can reach the standard, while the standard deviation of SBP can reach the standard by 1.84mmHg. At the same time, the results of the Bland-Altman analysis in this paper show that the algorithm in this paper can replace the PPG2ABP method. All the results show that the algorithm in this paper contributes a certain strength to the development of non-invasive blood pressure prediction methods.

This work proposes a non-invasive blood pressure prediction method based on sub-network aggregation and large convolution kernel neural network. The backbone network in this work integrates multi-scale information through maximum pooling, up-sampling, and layer-jumping connections; while the sub-network provides effective information to the backbone network, its parallel aggregation with the backbone network speeds up the blood pressure prediction. The superiority of this method compared to other methods that predicted ABP data for non-invasive blood pressure through PPG data has been verified on the MIMIC II data set. The algorithm in this work has potential to provide high-quality blood pressure data in wearable devices. How to reduce the size of the model while improve the accuracy of the algorithm is the future work.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 62072002, 62172004, 61872004, and U19A2064), Educational Commission of Anhui Province (No. KJ2019ZD05), and Anhui Scientific Research Foundation for Returness.

References

1. Roth, G.A., Mensah, G.A., Johnson, Z., et al.: Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J. Am. Coll. Cardiol.* **76**(25), 2982–3021 (2020)
2. Yan, W.R., Peng, R., Zhang, Y.T., et al.: Cuffless continuous blood pressure estimation from pulse morphology of photoplethysmograms. *IEEE Access* **99**, 141970–141977 (2019)
3. Chan, K.W., Hung, K., Zhang, Y.T.: Noninvasive and cuffless measurements of blood pressure for telemedicine. In: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3592–3593. IEEE, Istanbul, Turkey (2001)

4. Poon, C., Zhang, Y.T.: Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 5877–5880. IEEE, Shanghai, China (2005)
5. Yan, C., Wen, C., Tao, G., et al.: Continuous and noninvasive blood pressure measurement: a novel modeling methodology of the relationship between blood pressure and pulse wave velocity. *Ann. Biomed. Eng.* **37**(11), 2222–2233 (2009)
6. Monika, S., Martin, G., Matja, G., et al.: Non-invasive blood pressure estimation from ECG using machine learning techniques. *Sensors* **18**(4), 1160–1179 (2018)
7. Mousavi, S.S., Hemmati, M., Charmi, M., et al.: Cuff-less blood pressure estimation using only the ECG signal in frequency domain. In: International Conference on Computer and Knowledge Engineering. pp. 147–152. Department of Biomedical Engineering, Department of Electrical Engineering, University of Zanjan, Zanjan, Iran (2018)
8. Suzuki, S., Oguri, K.: Cuffless blood pressure estimation by error-correcting output coding method based on an aggregation of AdaBoost with a photoplethysmograph sensor. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6765–6768. IEEE, Hilton Minneapolis, MI, USA (2009)
9. El-Hajj, C., Kyriacou, P.A.: A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure. *Biomed. Signal Process. Control* **58**(9859), 101870 (2020)
10. Xing, X., Sun, M.: Optical blood pressure estimation with photoplethysmography and FFT-based neural networks. *Biomed. Opt. Express* **7**(8), 3007–3020 (2016)
11. Gao, S.C., Wittek, P., Zhao, L., et al.: Data-driven estimation of blood pressure using photoplethysmographic signals. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 766–769. IEEE, Orlando, FL, USA (2016)
12. Fujita, D., Suzuki, A., Ryu, K.: PPG-based systolic blood pressure estimation method using PLS and level-crossing feature. *Appl. Sci.* **9**(2), 304 (2019)
13. Brophy, E., Vos, M., Boylan, G., et al.: Estimation of Continuous Blood Pressure from PPG via a Federated Learning Approach [EB/OL] (2021). <https://arxiv.org/abs/2102.12245>
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. IEEE, Las Vegas, NV, USA (2016)
16. Szegedy, C., Liu, W., Jia, Y.Q., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE, Boston, MA, USA (2015)
17. Zhang, X.Y., Zhou, X.Y., Lin, M.X., et al.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 6848–6856. IEEE, US (2017)
18. Saeed, M., Villarroel, M., Reisner, A.T., et al.: Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. *Crit. Care Med.* **39**(5), 952–960 (2011)
19. Kachuee, M., Kiani, M.M., Mohammadzade, H., et al.: Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Trans. Biomed. Eng.* **64**(4), 859–869 (2016)
20. Wang, J., Sun, K., Cheng, T., et al.: Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3349–3364 (2021)

21. Ibtehaz, N., Rahman, M.S.: PPG2ABP: Translating Photoplethysmogram (PPG) Signals to Arterial Blood Pressure (ABP) Waveforms using Fully Convolutional Neural Networks [EB/OL] (2021). <https://arxiv.org/abs/2005.01669>
22. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. *Int. J. Nurs. Stud.* **47**(8), 931–936 (2010)



Integrating Knowledge Graph and Bi-LSTM for Drug-Drug Interaction Predication

Shanwen Zhang, Changqing Yu^(✉), and Cong Xu

College of Electronic Information, Xijing University, Xi'an 710123, China
zhangshanwen@xijing.edu.cn, wjdw716@163.com

Abstract. To solve the problem of ADDI extraction, the construction and implementation of knowledge graph provide a suitable solution for medical knowledge storage and management. This paper designs the construction and implementation of a knowledge graph based on deep learning. Named entity recognition and relationship extraction are carried out on the text of social media data, and then the graph database is used to store medical knowledge and construct the knowledge graph (KG). A DDI prediction method is proposed by combining knowledge graph (KG) and Bi-directional long-short-term memory network (Bi-LSTM) with attention. The multiple DDI sources are integrated by KG, and then are transformed into vectors by the knowledge representation model Hole. Finally, the implicit features of DDI are extracted by Bi-LSTM, and DDI is identified by Softmax classifier. The proposed method effectively combines the advantages of KG, BI-LSTM and attention mechanism, which can not only extract the global information of DDI, but also extract the sequence information of DDI. The experimental results on the DDI corpus dataset validate that the proposed method is effective and feasible for DDI prediction even with the reasonable architecture.

Keywords: Drug-drug interaction (DDI) · Adverse DDI (ADDI) · Knowledge graph (KG) · Bidirectional long-short-term memory network (Bi-LSTM)

1 Introduction

With the rapid development of related technology in the medical field, people pay more and more attention to the medical and health problems. There are more and more online medical and health websites on the Internet, and there are more and more ways for people to seek medical advice. People are now used to talking about illnesses and their treatments online and there is a lot of knowledge about drug interactions. However, these corpora belong to unstructured data, and there are no unified annotated corpora and annotated standards. Therefore, it has become a difficulty in medical research to construct medical corpora by combining their unique textual and structural features. Therefore, it is difficult to apply the model methods of entity recognition and relation extraction in the traditional domain to the network corpus, which brings great challenges to the ADDI extraction task based on natural language processing. ADDI will lead to fever, vomiting and other symptoms of the patient and increase the cost of treatment.

In serious cases, it will even cause a great threat to the patient's health and even death. Therefore, understanding the interaction between drugs and drugs is of great significance and value to the diagnosis and treatment of patients and the development of drugs. It is known that reasonable drug combination can enhance the efficacy or avoid adverse DDI (ADDI), while unreasonable drug combination may worsen a patient condition or even lead to increased death [1–4]. It is an important and challenging research due to the rich DDI information growing exponentially [5–8]. Zhu et al. [9] proposed a multi-task multi-attribute learning model (MTMAL) for ADDI prediction. In MTMAL, two drug attributes, molecular structure and side effect are utilized to model ADDI to uncover the adverse mechanisms among drugs. Cheng et al. [2] proposed a heterogeneous network-assisted inference framework to predict DDI by drug-drug pair similarities, including four features: phenotypic similarity, therapeutic similarity, chemical structural similarity and genomic similarity. Zhang et al. [10] proposed a measure of drug-drug similarity calculated in a drug feature space by exploring linear neighborhood relationship, then transferred the similarity from the feature space into the side effect space, finally predicted DDI by propagating the known side effect information through a similarity-based graph. Rohani et al. [11] predicted DDI by integrating similarity-constrained matrix factorization (ISCMF), where 8 kinds of similarities are calculated based on the drug substructure, targets, side effects, off-label side effects, transporters, pathways, enzymes, indication data and Gaussian interaction profile for the drug pairs.

Deep learning approaches avoid the designed feature extraction and can extract semantic features automatically for DDI prediction [12–15]. The limitation of CNN-based methods is that they suffer from the over-fitting issue, and neglect long distance dependency between the words in the candidate DDI instances, which may be helpful for DDI prediction. To incorporate the long distance dependency, Liu et al. [16] proposed a dependency-based CNN for DDI prediction, designed a simple rule to combine CNN with the CNN model to reduce error propagation. The shortest dependency path (SDP) between two entities contains valuable syntactic and semantic information. Inspired by the deep learning approaches in natural language processing, Yi et al. [17] proposed a recurrent neural network (RNN) with multiple attention layers for DDI prediction, and evaluated the model on 2013 DDIExtraction dataset. Shukla et al. [18] presented an integrated convolutional mixture density RNN by integrating CNN, RNN and mixture density networks. The extensive comparative analysis reveals that the proposed model significantly outperforms the competitive models. Park et al. [19] proposed an attention-based graph convolutional networks (AGCN) for DDI prediction, where AGCN is designed to leverage contextual and structural knowledge together in combination with encoders based on recurrent networks.

Through knowledge graph (KG), the information of the Internet can be not only expressed in a form that more closely resembles the human cognitive world, but also provides a better way to organize, manage and utilize the vast amount of information [20]. To alleviate the deficiencies in ADDI extraction, KG is used to describe the logical association between drugs, and between drug and its treatment, and to extract DDIs. Medical KG can combine the doctor professional knowledge and thinking mode, which contains knowledge content, knowledge quantity and professional logic correlation between knowledge. It can be used to compute several similarity measures between

all the drugs in a scalable and distributed framework [21, 22]. Lin et al. [23] predicted DDI by KG neural network (KGNN), evaluated DDI with respect to the pharmacokinetics and safety of dotinurad when co-administered with oxaprozin, and compared its pharmacokinetic parameters and evaluated safety. HolE is a compositional vector space based model for KG [24]. It combines the expressive power of the tensor product with the efficiency and simplicity of TransE. The representation of all entities and relations are learned jointly by HolE. HolE can handle relatively large KG and provide better performance compared to the state-of-the-art embedding techniques [25]. So it can propagate the information between triples which can capture the global dependency in the data. Attention mechanism is now widely applied to many areas such as machine translation, voice recognition, image tracking, and can improve the performance of RNN and LSTM [26].

There is a lot of DDI information which can be integrated by KG. The integrated DDI information by KG is useful for developing high-accuracy extraction model of DDI, and Bi-LSTM is good at accumulating feature information in both the forward and backward directions. A DDI prediction method is proposed by combining KG, Bi-LSTM and attention mechanism, consisting of preprocessing the DDI information from the corpus dataset, constructing a KG of DDI, transforming the DDI information from KG to vectors suitable to the identification model, finally presenting a modified Bi-LSTM with attention to identify DDI.

2 Drug-Drug Interaction Predication Method

A DDI prediction method by combining KG and Bi-LSTM is proposed, and its architecture is shown in Fig. 2. In DDI named entity recognition method, KG is used for DDI information integration, Bi-LSTM is used for feature extraction, small sample annotation data set is used to train the entity recognition model, extract the linguistic features and structural features of DDI corpus, then continuously amplify the annotation dataset, and repeatedly iterate and optimize the DDI predication model. Bi-LSTM can solve the problem of gradient disappearance and long-term dependence of RNN model by controlling the “forgetting gate”, “input gate” and “output gate” of traditional RNN model. In view of the feature selection problem of relation extraction between entities, attention is introduced into Bi-LSTM, and attention layer is added after Bi-LSTM layer for classification of relations between entities. In the attention layer, the sentence-level weight vector is generated through model training. In the model test, the input vector of the attention layer is multiplied by the weight vector, and the feature vector of the word level is converted into the feature vector of the sentence level, so as to reduce information redundancy and information loss in the process of feature extraction.

KG is denoted as $G_{KG} = (E \cup \phi(E), R \cup \phi(R))$, where $E = \{e_1, e_2, \dots, e_N\}$ contains N entities, $R = \{r_1, r_2, \dots, r_M\}$ has M entity relations, $T = \{t_1, t_2, \dots, t_5\}$ is DDI type set. Each entity e or DDI r is mapped to one of five semantic types by a relational mapping function $\phi(e) \rightarrow T_e$. Given a path $\pi_i^l = e_0 r_0 e_1 r_1 \dots r_{l-1} e_l$ in KG expressed a sequence of continuous drugs and their DDIs, where e_i is the i -th drug, r is the i -th

DDI type. Any DDI as a triple ($drug_i$, DDI_i , $drug_{i+1}$) is obtained by the path search algorithm, defined as follows,

$$\pi^l = \rho(drug_i \rightarrow drug_{i+1}; DDI_i, l) \quad (1)$$

where π^l represents all the paths in KG with a starting point $drug_i$ and an ending point $drug_{i+1}$, and the length through the node is DDI_i .

A path set $l = \{\pi^2, \pi^3, \dots, \pi^l\}$ with length 2 to l is further constructed as the positive example set of this experiment. Similarly, the set of negative example set in the experiment is obtained by triples $(drug'_j, DDI'_j, drug'_{j+1})$, indicating that two drugs have no DDI.

The goal of DDI prediction model is to evaluate the probability of DDI type between two drugs. DDI prediction can be obtained by the KG embedding model HolE [25], defined as follows,

$$\Pr(\phi_r(h, t) = 1 | \Phi) = \sigma(r^T (e_h \odot e_t)) \quad (2)$$

where r , e_h , e_t are vector representations of the relations and entities, $\sigma(x) = 1/[1 + \exp(-x)]$ is the logistic function, $\Phi = \{e_i\}_{i=1}^{n_e} \cup \{r_k\}_{i=1}^{n_r}$ is the set of all embedding, \odot denotes the compositional operator to create a composite vector representation for the pair (s, o) from the embedding e_h , e_t .

Bi-LSTM is used to design a DDI prediction model. Its two hidden layers process the data in different directions,

$$\begin{aligned} h_{ft} &= H(W_{xh_f}x_t + W_{h_fh_f}h_{f_{t-1}} + b_{h_f}), \\ h_{bt} &= H(W_{xh_b}x_t + W_{h_bh_b}h_{b_{t-1}} + b_{h_b}), \end{aligned} \quad (3)$$

where h_f and h_b are the hidden layer of the forward and back layers, respectively.

Then the probability of DDI in the input model is

$$p(y|X) = \sigma(W_{h_fz}h_t + W_{h_bz}h_t + b_z) \quad (4)$$

where W_{h_fz} , W_{h_bz} and b_z are the parameters to be trained.

In order to prevent the training mode from over-fitting, dropout is added in the non-cyclic part of Bi-LSTM. Time is used to optimize the Cross Entropy Loss Function $L(\theta)$,

$$L(\theta) = -\frac{1}{n} \sum_{i=1}^n (y \ln(p(y|X_i)) + (1-y) \ln(1-p(y|X_i))) \quad (5)$$

Attention mechanism emphasizes the importance of physical information in sentences to improve the performance of Bi-LSTM [26]. Given a drug and its corresponding pharmacological features, drug class features and drug textual description features, attention mechanism is applied to assign different weights according to the specific role that each feature interacts with other features. The representation of drug class feature v_c is defined as:

$$v_c = H_w \alpha_w^T \quad (6)$$

where $M_w = \tanh(W_{sw}H_w)$, $\alpha_w = \text{softmax}(w_w^T M_w)$, $M_w \in R^{dl \times m}$ is a nonlinear mapping function, $W_{sw} \in R^{dl \times dl}$ and $w_w \in R^{dl}$ are projection parameters, $\alpha_w \in R^m$ is the normalized attention. Other two types of features are processed by the same attention mechanism. Then these three feature embedding types are concatenated for the final DDI representation.

For DDI prediction, a joint layer is added to join the final drug representations of drug1 and drug2. The outputs of the Bi-LSTM and fully connected layer then go through a Softmax layer for binary classification:

$$y = \text{softmax}(w_0 v_c + b_0) \quad (7)$$

where each dimension of y is the normalized probability of a certain relation, i.e., positive correlation or negative correlation, in accordance with the fully connected layer, $w_0 \in R^{2 \times dl}$ is the projection matrix, and $b_0 \in R^2$ is the offset vector.

Extracting DDI in the text is the process of identifying how two target drugs in a given sentence interact. Figure 1 shows the vector transformation process based on KG jointing and embedding.

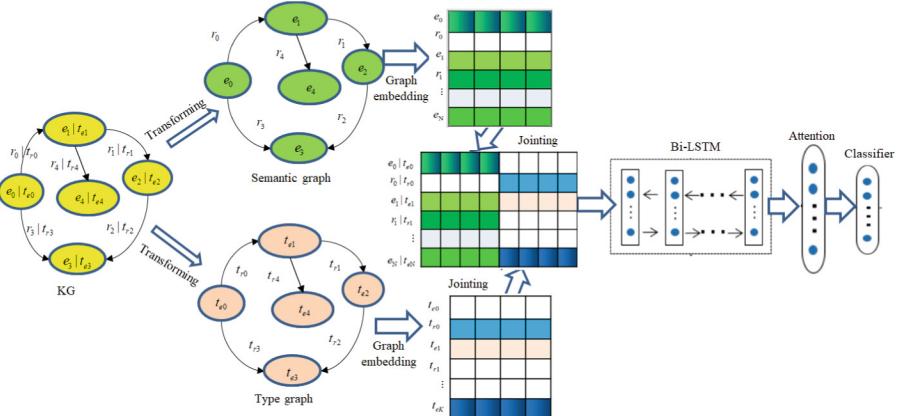


Fig. 1. Vector transformation process based on KG jointing and embedding

From the above analysis, we give the process of the proposed DDI prediction method based KG and Bi-LSTM with attention, as shown in Fig. 2.

At last, the Softmax classifier is used to obtain a normalized probability score for each class. *Recall*, *Precision* and *F-Score* metrics are often used on testing set to evaluate DDI prediction performance, and C denotes the set of {Advice, Effect, Int, Mechanism and Negative}. The precision and recall of each $c \in C$ are calculated by, defined as follows,

$$P_c = \frac{\text{The number of that DDI is } c \text{ and classified as } c}{\text{The number of that DDI is classified as } c} \quad (8)$$

$$R_c = \frac{\text{The number of that DDI is } c \text{ and classified as } c}{\text{The number of that DDI is } c} \quad (9)$$

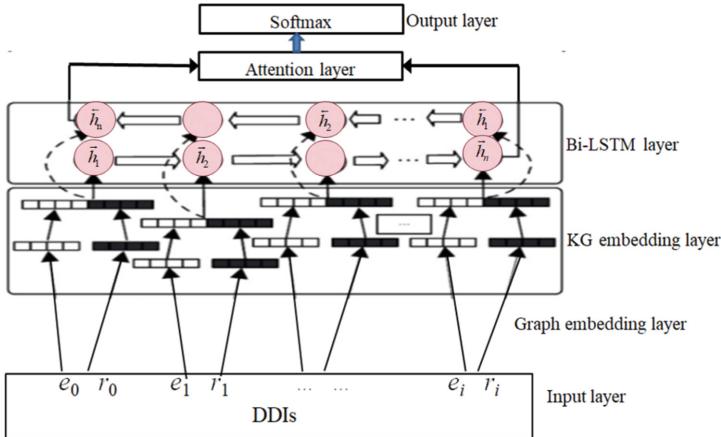


Fig. 2. KG identification process

Then the overall *Precision*, *Recall* and *F-Score* are calculated as,

$$\text{Precision} = \frac{1}{|C|} \sum_{c \in C} P_c, \quad \text{Recall} = \frac{1}{|C|} \sum_{c \in C} R_c, \quad \text{F-score} = \frac{2PR}{P+R} \quad (10)$$

3 Experiments and Analysis

There are a large number of drugs introduced every year and a number of DDIs also has quick growth. A freely available web resource for DDI prediction and identification was developed (<http://way2drug.com/ddi/>). DrugBank is a bioinformatics and chemical informatics database provided by the University of Alberta. A sentence with two drug names represents a DDI instance, which is used as the input to the proposed model. The DDI corpus dataset is a benchmark dataset for identifying DDI [24]. It contains a total of 1025 documents, 233 Medline abstracts (DDI-MedLine) and 792 texts from the DrugBank database (DDI-DrugBank), which were manually annotated with 18,502 drugs and 5028 DDIs. Some examples of the DDI corpus in brat format can be found at <http://brat.nlplab.org/>. The pre-training corpus data for word representations is about 2.5 gigabytes in size, consisting of two parts, i.e., abstracts and texts, which can be used to extract the type features and semantic features. The DDI corpus dataset is unbalanced, i.e., there are more Effect class and Mechanism instances in the positive case, while Int instances occupy only a small part. So, down-sampling is used for the Effect class and the Mechanism class, and up-sampling the Int class to handle the problem of class imbalance. For simplicity, in positive dataset, all drug pairs in each sentence are annotated manually, and each pair is annotated as either no interaction or true interaction with more fine-gained annotations consisting of four labels: Advice, Effect, Int, Mechanism (the classification task). The rest are annotated as negative instances. The 5-fold cross validation scheme is adopted to implement the experiments, i.e., 80% of the data are used for training all parameters of KG and Bi-LSTM, and the optimized model is evaluated on

20% held-out data in which the best hyper-parameters are produced through a random search.

The proposed model is to train and test on the DDI corpus dataset and compared with four methods based on feature extraction (FE) [4], CNN [12], RNN [17], Knowledge Graph Embedding and Convolutional-LSTM Network (KGC-LSTM) [22]. The experiments are conducted on 32G memory, with Intel Core i5-4200U CPU @2.30 GHz, GPU GEFORCE GTX 1080ti, Ubuntu14.0. The deep learning architecture is Tensorflow1.7.0, Keras, and LSTM. Bi-LSTM is optimized using Adam Optimizer with learning rate 0.01 and mini batch 1,000, all other parameters are randomly initialized from $[-0.1, 0.1]$, the maximum length of sentence is set to 100.

The 5-fold cross validation experiment is repeated 50 times, and their averages are as the DDI prediction results. The results of the proposed method and four comparative methods on the test set while training are conducted using either the complete and filtered dataset are shown in Table 1.

Table 1. The results of 5 methods on the complete and filtered datasets

Methods	Precision (%)	Recall (%)	F-score (%)
FE	63.73	62.37	63.04
CNN	67.25	65.91	66.57
RNN	72.44	67.52	69.89
KGC-LSTM	73.63	71.32	72.46
Our method	74.25	72.14	73.18

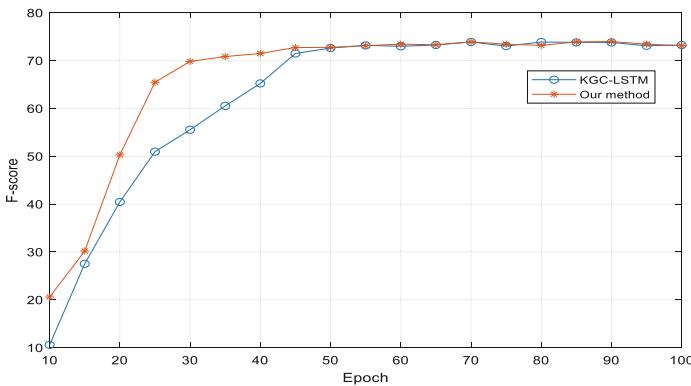


Fig. 3. F-scores of KGC-LSTM and our method versus iterations

To demonstrate performance of the proposed method, Fig. 3 shows the F-scores of KGC-LSTM and our method versus the iterations. From Fig. 3, it is found that

our method is better than KGC-LSTM on the whole. Before 50 Epochs, our method is superior to KGC-LSTM obviously. After 50 Epochs, the training process of two methods is relatively stable. The results show that the convergence of our method is better than KGC-LSTM. The main reason is that the intensive module and attention mechanism are utilized in our method.

4 Conclusions

DDI prediction is an important and challenging research. Because the DDIs information in the corpus dataset is large and complicated, using the traditional rule-based, similarity-based and feature-based methods cannot effectively implement semantic understanding and feature extraction of the DDIs. KG-based approach using multiple data sources is comparable to current state-of-the-art methods. Bi-LSTM is often utilized to learning the identification features to complete DDI prediction. DrugBank consists of manually curated texts, collected from various sources and verified by accredited experts. In the paper, a DDI prediction method is proposed by integrating KG, Bi-LEST and attention mechanism, and is validated on the public dataset. The results show that the proposed method is effective to extract DDI. In the future, we will further improve overall representational learning, and consider that instance generation using generative adversarial networks would cover the instance shortage in specific category.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Nos. 62172338 and 62072378).

References

1. Sandson, N.B., Armstrong, S.C., Cozza, K.L.: An overview of psychotropic Drug-Drug interactions. *Psychosomatics* **46**(5) (2005). <https://doi.org/10.1176/appi.psy.46.5.464>
2. Cheng, F., Zhao, Z.: Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J. Am. Med. Inf. Assoc. Jamia* **e2**, 278–286 (2014)
3. Hina, H., Huma, A., Farya, Z., et al.: Drug-Drug interaction. *Profess. Med. J.* **24**(3) (2017). <https://doi.org/10.17957/TPMJ/17.3670>
4. Raihani, A., Laachfoubi, N.: Extracting Drug-Drug interactions from biomedical text using a feature-based kernel approach. *J. Theor. Appl. Inf. Technol.* **92**, 109–120 (2016)
5. Cami, A., Manzi, S., Arnold, A., et al.: Pharmacointeraction network models predict unknown drug-drug interactions. *PLoS ONE* **8**(4), e61468 (2013)
6. Bui, Q.C., Sloot, P.M., Mulligen, E.M., et al.: A novel feature-based approach to extract Drug-Drug interactions from biomedical text. *Bioinformatics* **2014**(23), 3365–3371 (2014)
7. Kim, S., Liu, H., Yeganova, L., et al.: Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. *J. Biomed. Inform.* **55**, 23–30 (2015)
8. Jamal, S., Goyal, S., Shanker, A., et al.: Predicting neurological adverse Drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Rep* **7**(1), 872 (2017). <https://doi.org/10.1038/s41598-017-00908-z>
9. Zhu, J., Liu, Y., Wen, C.: MTMA: multi-task multi-attribute learning for the prediction of adverse Drug-Drug interaction. *Knowl. Based Syst.* **199**, 105978 (2020). <https://doi.org/10.1016/j.knosys.2020.105978>

10. Zhang, W., Yue, X., Liu, F., et al.: A unified frame of predicting side effects of drugs by using linear neighborhood similarity. *BMC Syst. Biol.* **11**(S6), 101 (2017). <https://doi.org/10.1186/s12918-017-0477-2>
11. Rohani, N., Eslahchi, C., Katanforoush, A.: ISCMF: integrated similarity-constrained matrix factorization for Drug–Drug interaction prediction. *Network Model. Anal. Health Inf. Bioinf.* **9**(1), 1–8 (2020). <https://doi.org/10.1007/s13721-019-0215-3>
12. Liu, S.Y., Tang, B.Z., Chen, Q.C., et al.: Drug–Drug interaction extraction via convolutional neural networks. *Comput. Math. Methods Med.* **2016**, 6918381 (2016)
13. Sun, X., Feng, J., Ma, L., et al.: Deep convolution neural networks for Drug–Drug interaction extraction. *IEEE Int. Conf. Bioinf. Biomed.* (2018). <https://doi.org/10.17816/PAVLOVJ2013370-76>
14. Zhao, Z.H., Yang Z.H.L., et al.: Drug–Drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* **32**(22), 3444–3453 (2016)
15. Víctor, S.P., Isabel, S.B.: Evaluation of pooling operations in convolutional architectures for Drug–Drug interaction. *BMC Bioinf.* **19**(Suppl 8), 209 (2018)
16. Liu, S., Chen, K., Chen, Q., et al.: Dependency-based convolutional neural network for drug-drug interaction extraction. *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1074–1080 (2017)
17. Yi, Z., et al.: Drug–Drug interaction extraction via recurrent neural network with multiple attention layers. In: Cong, G., Peng, W.-C., Zhang, W.E., Li, C., Sun, A. (eds.) *ADMA 2017. LNCS (LNAI)*, vol. 10604, pp. 554–566. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69179-4_39
18. Shukla, P.K., Shukla, P.K., Sharma, P., et al.: Efficient prediction of Drug–Drug interaction using deep learning models. *IET Syst. Biol.* **14**(4), 211–216 (2020)
19. Park, C., Park, J., Park, S.: AGCN: attention-based graph convolutional networks for Drug–Drug interaction extraction. *Expert Syst. Appl.* **159**, 113538 (2020)
20. Abdelaziz, I., Fokoue, A., Hassanzadeh, O., et al.: Large-scale structural and textual similarity-based mining of knowledge graph to predict Drug–Drug interactions. *J. Web Seman.* **44**, 104–117 (2017)
21. Shen, Y., et al.: KMR: knowledge-oriented medicine representation learning for Drug–Drug interaction and similarity computation. *J. Cheminform.* **11**(1), 1–16 (2019). <https://doi.org/10.1186/s13321-019-0342-y>
22. Karim, M.R., Cochez, M., Jares, J.B., et al.: Drug–Drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. *ACM-BCB 2019*, pp. 113–123. Niagara Falls, USA (2019)
23. Lin, X., Quan, Z., Wang, Z.J., et al.: KGNN: knowledge graph neural network for Drug–Drug interaction prediction. In: *29th International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence* (2020). <https://doi.org/10.24963/ijcai.2020/376>
24. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. *Computer Science* (2015). [arXiv:1510.04935, https://arxiv.org/pdf/1510.04935.pdf](https://arxiv.org/pdf/1510.04935.pdf)
25. Wu, H., Xing, Y., Ge, W., et al.: Drug–drug interaction extraction via hybrid neural networks on biomedical literature. *J. Biomed. Inform.* **106**, 103432 (2020)
26. Li, X., Zhang, W., Ding, Q.: Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Sig. Process.* **161**, 136–154 (2019)



A 3D Medical Image Segmentation Framework Fusing Convolution and Transformer Features

Fazhan Zhu¹ , Jiaxing Lv¹ , Kun Lu^{1,2} , Wenyan Wang^{1,2,3} ,
Hongshou Cong¹ , Jun Zhang⁴ , Peng Chen⁵ , Yuan Zhao^{1,2} ,
and Ziheng Wu^{1,2()}

¹ School of Electrical and Information Engineering, Anhui University of Technology,
Ma'anshan 243032, Anhui, China
wziheng@ahut.edu.cn

² Key Laboratory of Metallurgical Emission Reduction and Resources Recycling, (Anhui
University of Technology), Ministry of Education, Ma'anshan 243002, China

³ School of Materials Science and Engineering, Anhui University of Technology,
Ma'anshan 243032, Anhui, China

⁴ School of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui,
China

⁵ National Engineering Research Center for Agro-Ecological Big Data Analysis and
Application, School of Internet and Institutes of Physical Science and Information Technology,
Anhui University, Hefei 230601, Anhui, China

Abstract. Medical images can be accurately segmented to provide reliable basis for clinical diagnosis and pathology research, and assist doctors to make more accurate diagnosis, as well as deep learning technology can accelerate this process. Convolutional Neural Networks (CNNs) and Transformer have become two mainstream architectures of deep learning in medical image segmentation. However, the Transformer architecture has limited ability to obtain local inductive bias, and the Transformer architecture is at a disadvantage in a small sample data set. Many theories and experiments show that the above problems can be effectively solved by fusing Convolution and Transformer features. In this manuscript, a new U-shaped segmentation model based on Convolution and swin-transformer framework is proposed, which is called CST-UNET. In the encoder part, it combines the advantages of both dilated convolution and Transformer, which can make the model fully obtain semantic inductive bias information and long-term information. At the same time, it has the advantages of fewer parameters and lower Flops. Even if it is trained on a small sample data set, the framework still has strong generalization ability. In addition, on BraTS2021 dataset, the Dice coefficients of ET, TC and WT are 85.46%, 89.38%, 92.35% respectively, and the result of HD95 are 7.95, 5.06 and 4.07 respectively.

Keywords: Medical image segmentation · 3D · Convolutional neural networks · Transformer

1 Introduction

Automatic and accurate segmentation of these malignant tumors by magnetic resonance imaging (MRI) is crucial to clinical diagnosis [4]. UNET [5] is the first U-shaped segmentation network completely based on convolutional architecture. In the field of medical image segmentation, its design concept of encoder-decoder has been used for reference by many researchers and it is widely used. The target of image segmentation is usually multi-scale [6]. The previous image segmentation models, FCN [7] and SegNet [8], used pooling layer, which is not conducive to segmentation of the multi-scale target. In our segmentation framework, pooling layer is not used, but dilated convolution is adopted at each stage of the convolution branch of the encoder part. On the one hand, when features are extracted from images of the same size, the dilated convolution of large convolution kernels can reduce the number of model parameters, on the other hand increase the model's receptive field [6]. Furthermore, Group-Norm [9] is used in the framework instead of Batch-Norm [10]. The reason is that our framework directly processes 3D medical images, which consumes a lot of GPU memory. Due to the limitation of GPU graphics memory capacity, the batch size of tensor is set to 2 at most, which leads to the limitation of Batch-Norm normalization ability. However, Group-Norm will break through the limitation. In addition, we also changed the ReLU activation function to SiLU [11], which can improve the robustness of the model segmentation.

The main idea of ResNet [12], the first residual network, is to add a direct channel in the framework, that is, the idea of Highway Network. It effectively solves the problems of information loss, gradient disappearance or gradient guarantee in traditional convolution networks or fully connected networks when information is transmitted. Therefore, the residual connection is also used in the nonlinear mapping layer of our encoder part of the framework.

Window self-attention put forward by Swin Transformer [13], which effectively solves the problem of high computational complexity of self-attention in Vision Transformer (ViT) [14], and the strategy of shift windows subtly links the information interaction between windows.

Conformer [15], a hybrid network structure, takes advantage of convolutional operations and self-attention mechanisms for enhanced representation learning. TransUNet [16], the first 2D medical image segmentation framework, employs a hybrid CNN-Transformer architecture to leverage both detailed high-resolution spatial information from CNN features and the global context encoded by Transformer. However, they are both 2D architectures. Besides, TransUNet only utilizes convolutional layer for feature extraction in the first layer of encoder, and only makes use of the deconvolution layer for up-sampling in the decoder part, which results in the framework not fully reflect the advantages of the combination of convolution and Transformer. UNETR [17] applies ViT to volume medical image segmentation, but it contains 12-layer Transformer, resulting in parameter redundancy of the model and increasing training difficulty, so the generalization ability of the model is limited. nnFormer [18] puts forward a strategy of transferring Query, Key, Value between encoder and decoder, i.e., skip attention. VT-UNet [19] is a volume medical image segmentation network completely based on swin-transformer. On the basis of nnFormer, an independent learning swin-transformer branch is added to the decoder part, and the two swin-transformer branches are linearly fused, and 3D

Fourier position coding is added in VT-UNet. VT-UNet exceeds the previous framework in segmentation accuracy. Due to the less induction bias injected into the model during encoding, the recognition of the target boundary needs to be improved in segmentation. The high HD95 value segmented on the BraTS2021 data set indicates that it has not learned enough segmentation target boundary features [20–22]. Compared with ViT, the window self-attention and shifting window strategy [13] greatly reduce the Flops of the model, but to some extent, it also limits the ability of the model to obtain long-term related information. In order to solve the above problems, we propose a new 3D feature fusion structure based on dilated convolution with a larger kernel and swin-transformer.

In conclusion, our contributions are as follows: (1) A new volume pixel segmentation framework combining convolutional and Transformer features is proposed; (2) A feature fusion method with learnable weight weighting method is designed; (3) A dynamic weighted loss function is designed; (4) Compared with the previous methods, the segmentation results on the BraTS2021 data set are improved.

2 Relative Work

2.1 CNN Based Methods

Deep learning based on convolutional neural network is widely used in image segmentation [23–25]. UNET [5] is the first to propose a U-shaped encoder-decoder network to segment cell images, and there are many subsequent variations [26–28]. Some methods that can directly segment volume pixels have been proposed in the decades of rapid development of deep learning [29, 30]. The design idea of residual connection was proposed by ResNet [12], which effectively compensates for the loss of deep network and reduces the risk of gradient disappearance or explosion in network training process.

2.2 Transformer Based Methods

Transformer [31] is a great reminder of the NLP field, The intention of Vision Transformer (ViT) [14] is to break through the barriers of text processing and image processing, by simply changing Transformer, and cleverly uses Transformer structure in the image field. There are many subsequent explorations of Transformer [32–35]. The strategy of sliding window is proposed by swin-transformer [13], which successfully optimizes the problem of high computational complexity of ViT. nnFormer [18] and VT-UNET [19] are two 3D image segmentation architectures based entirely on Transformer architecture.

2.3 CNN and Transformer Based Methods

TransUNet [16] is the first proposed medical image segmentation framework combining convolution and Transformer [36–38], but it only uses a convolution layer to extract features for Transformer, and it is a 2D image segmentation network. ViTAE [39] has several spatial pyramid reduction modules to down-sample and embed the input image into tokens with rich multi-scale context by using multiple convolutions with different dilation rates. UNETR [17], TransBTS [4] and SWIN UNETR [40], are 3D medical

image segmentation architecture based on traditional convolution and Transformer, but they are all based on Transformer in conventional ViT, so the model still has the problem of high computational complexity. Therefore, we propose an architecture based on feature fusion of dilated convolution and swin-transformer. It can not only segment 3D medical images accurately, but also has fewer parameters and lower Flops.

3 Method

3.1 Overview of the Architecture

The overview of Convnet and Swin Transformer U-shaped Network (CST-UNET) is presented in Fig. 1, which mainly consists of three parts, i.e., the encoder, bottleneck and decoder. More specifically, the encoder part is composed of a Convolution branch, a Transformer branch and four feature fusion submodules for fusing Convolution and Transformer features, as well as the Convolution and Transformer branch are parallel, with four layers, not cross-connected. In the convolution part of encoder, each layer is made up of successively connected dilated convolution (DCNN), Group-Norm, SiLU and nonlinear projective layer. Like the backbone of swin-transformer [13], the first layer of Transformer in the encoder part is an embedding layer, and each subsequent stage consists of Transformer block (each block contains two successive layers) and 3D patch merging layer. However, there are two Transformer blocks in the last stage. The decoder part is completely composed of two swin-transformer branches, and the number of layers of these two branches corresponds to the encoder part. In order to restore the original image size, the 3D patch merging layer is replaced by 3D patch expanding. Besides, the bottleneck comprises a down-sampling layer and an up-sampling layer. Inspired by UNET, we add skip connections between corresponding feature pyramids of the encoder and decoder in a symmetrical manner, which helps to recover fine-grained details in the prediction. However, different from atypical skip connections that often use summation or concatenation operation, we introduce skip attention to bridge the gap between the encoder and decoder. The input data of CST-UNET is $x \in R^{H \times W \times D \times C}$, where H , W and D denote the height, width and depth of each input scan, respectively. C represents the number of medical image (e.g., MRI) in different modalities.

3.2 The CNNs Branch

As shown in Fig. 1(a), the convolution of the encoding part has four stages, namely, ET-stage0, ET-stage1, ET-stage2, ET-stage3. Transformer branch in encoder also has four stages in parallel with CNNs. Stacking mode of each layer's convolution block is the same in CNNs branch. The reason why the Group-Norm is adopted instead of the general BatchNorm is that, limited to the size of GPU memory, the BatchNorm size of medical images is usually very small, which is 2 in our experiment, and the BatchNorm cannot give full play to its normalization effect. GroupNorm [9] is a normalized layer designed for the case of small batch size. Then there is the SiLU activation function. Generally, SiLU is more robust than ReLU in image segmentation. At the end of each convolution, a nonlinear projection layer is added, and the structure of each nonlinear projection

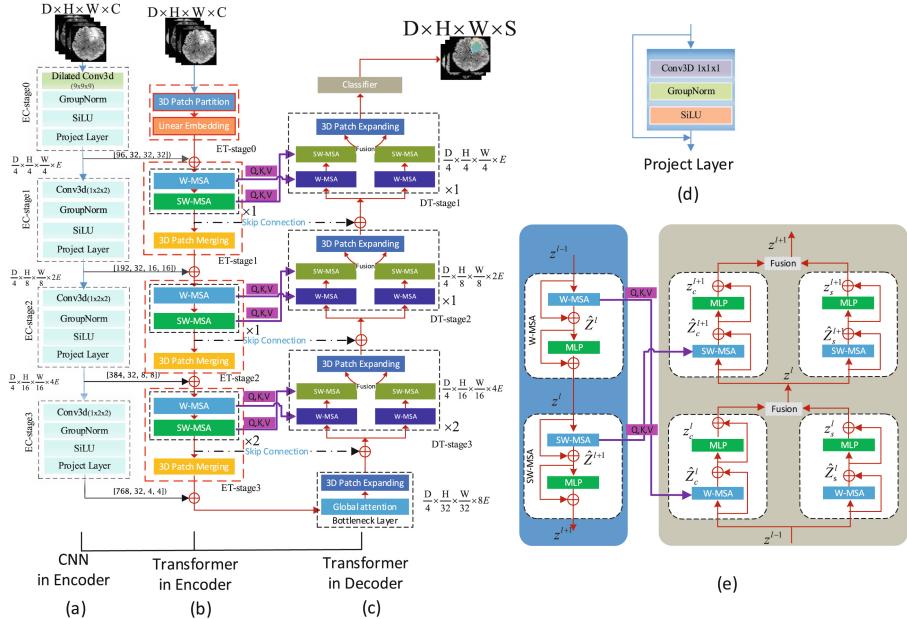


Fig. 1. The Overview of CST-UNET. (a) is the convolutional branch of the encoder part, with a total of 4 EC-stages separated by black dotted boxes. (b) illustrates the Transformer branch of the encoder part, with a total of four ET-stages separated by red dotted boxes. In addition, the blue solid line represents the transmission direction of information flow between the convolution blocks, and the black solid line represents the transmission path of features generated by the convolution branch to the Transformer branch. (c) shows the decoder part of this framework, which contains three DT-stages. Each DT-stage contains two parallel swin-transformer blocks and a 3D Patch Expanding layer, separated by black dotted boxes in the figure. (d) is the project layer in each ET-stage. (e) displays the way of skip attention between encoder and decoder. (Color figure online)

layer of each layer is the same: a 3-D convolution with a convolution kernel size of 1, GroupNorm, SiLU, and residual connection proposed by ResNet [12] is adopted, as shown in Fig. 1(d). Each stage contains two GroupNorm, and the number of groups at the same stage are equal, but the number of groups in different layers is different, which is the same as the number of heads of multi-head self-attention in the Transformer branch, which is 3, 6, 12, 24 in turn. In the first layer of convolution, we adopt the dilated convolution with convolution kernel of 9, expansion coefficient of 4 and step size of 3, which are the same in the three dimensions of $D \times H \times W$. In this way, the receptive field can be increased and make up for the ability of swin-transformer to obtain semantic information within the window (focusing too much on local information) and it is failure to fully obtain relevant global information. It can also supplement the local inductive bias of swin-transformer. The convolution kernel size of dilated convolution in the other three stage is $1 \times 2 \times 2$, expansion coefficient is $2 \times 1 \times 1$, step size is $1 \times 2 \times 2$, and the corresponding dimensions are Depth, Height and Weight respectively. Expansion coefficient is set to 2 in Depth dimension. For transferring a small amount of semantic

information between images of different depths to Transformer by convolution block, we conducted a subsampling operation, Patch Merging.

For the transformer part, we unified 32 in the depth direction. 32 images with continuous depth are selected sequentially for 3D down-sampling, and then the channel direction is spliced. The features of the convolution output of the previous encoder part will be fused with the features of the Transformer output, and the fused features will be fed into the next Transformer block, but not the convolution block, i.e., $z_l = \alpha \cdot x_{l-1} + \beta \cdot z_{l-1}$, where α, β are the two learnable weights, x, z represent the features of convolution and transformer output, respectively. After a 3D normal convolution layer operation, the size change of a tensor with size $D \times H \times W$ can be summarized as follows:

$$\left(\frac{D - F + 2P}{S} + 1 \right) \times \left(\frac{H - F + 2P}{S} + 1 \right) \times \left(\frac{W - F + 2P}{S} + 1 \right) \times C \quad (1)$$

where D, H and W are the depth, height and width of the image before the convolution operation, F is the side length of the convolution kernel, C is the number of the convolution kernel, S is the step size of the convolution operation, and P is the zero-padding size.

The calculation of dilated convolution is summarized as follows:

$$\left(\frac{D - F' + 2P}{S} + 1 \right) \times \left(\frac{H - F' + 2P}{S} + 1 \right) \times \left(\frac{W - F' + 2P}{S} + 1 \right) \times C \quad (2)$$

where $F' = d \times (F - 1) + 1$, d is a dilatation coefficient.

The calculation of a single convolution block in the encoder part is summarized as follows:

$$\begin{aligned} \hat{x}_0 &= DCNN(x_0) \\ \hat{x}_1 &= GN(\hat{x}_0) \\ \hat{x}_2 &= SiLU(\hat{x}_1) \\ x_1 &= PL(\hat{x}_2) + \hat{x}_2 \end{aligned} \quad (3)$$

where $DCNN$ denotes a dilated convolution neural network, GN is the GroupNorm, PL means that Projective Layer.

The calculation of a single project layer in the encoder part is summarized as follows:

$$\begin{aligned} \hat{x}_{2,1} &= Conv_{1 \times 1 \times 1}(\hat{x}_2) \\ \hat{x}_{2,2} &= GN(\hat{x}_{2,1}) \\ \hat{x}_{2,3} &= SiLU(\hat{x}_{2,2}) \end{aligned} \quad (4)$$

where $Conv_{1 \times 1 \times 1}$ denotes a 3D convolution with a convolution kernel of size $1 \times 1 \times 1$.

3.3 Transformer Branch

Swin-transformer [13], a variant of ViT [14], is used in the transformer branch of encoder and decoder. The transformer branch of the encoder part consists of four stages named

ET-Stage0, ET-Stage1, ET-Stage2, and ET-Stage3. In Fig. 1(b), each stage is separated by an independent red dotted box. In order to solve the problem that ViT directly calculates the self-attention of the whole image, which is too complicated, swin-transformer partitions the whole image into patches on average before calculating the self-attention, and then calculates the self-attention score on smaller patches. With each stage, the output tensor size is halved in the H and W dimensions, but stays the same in the D dimension, making it easier for Transformer to focus on calculating long correlations with deep semantic information. But in the convolution block of the encoder part, the model relates the information of pictures of different depths. The 3D patch partition in ET-Stage0 cuts the original image tensor to a quarter of the original depth, height and width, and the Linear Embedding maps each small image block after cutting to the E-dimensional space. $z_0 \in \mathbb{R}^{\frac{D}{4} \times \frac{H}{4} \times \frac{W}{4} \times E}$, z_0 represents the tensor after ET-stage0, D, H, W, E represent the dimensions of Depth, Height, Weight and Embedding dimension respectively. The next three stages are composed by swin-transformer blocks and a complex 3D patch. For merging 3D patch, the output dimension is $z_i \in \mathbb{R}^{\frac{D}{4} \times \frac{H}{(2^{i+2})} \times \frac{W}{(2^{i+2})} \times (2^{i \cdot E})}$, $i \in \{1, 2, 3\}$. But ET-Stage3 has two layers of swin-transformer, ET-Stage1 and ET-Stage2 both have only one layer. With the deepening of spatial dimension of mapping, the long-term correlation between semantics requires more parameters to learn. Therefore, we increase the number of layers of swin-transformer in ET-stage 3.

The bottleneck layer is located at the bottom of Fig. 1(c), which consists of a global attention layer together with a 3D Patch Expanding layer. This decoder part has two transformer branches, and the design idea of the encoder part transformer is roughly the same. As shown in Fig. 1(c), the decoder part has three stages, namely DT-stage1, DT-stage2 and DT-stage3. Each stage is composed of swin-transformer block and a 3D patch expanding layer. The 3D patch expanding layer can be regarded as the inverse process of the 3D patch merging layer. From DT-stage3 to DT-stage1, the number of layers of swin-transformer blocks in each stage is 2, 1, 1 in turn. The final layer of the decoder is the classification layer. In short, each swin-transformer block can be represented as follows:

$$\begin{aligned} \hat{z}^l &= W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \\ z^l &= MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \\ \hat{z}^{l+1} &= SW - MSA\left(LN\left(z^l\right)\right) + z^l \\ z^{l+1} &= MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \end{aligned} \quad (5)$$

where W -MSA and SW -MSA denote regular and window partitioning self-attention modules, respectively. \hat{z}^l and \hat{z}^{l+1} are the outputs of W -MSA and SW -MSA; LN and MLP denote layer normalization and Multi-Layer Perceptron. Self-attention calculation formula is as follows:

$$SA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{E}} + \mathbf{B}\right)\mathbf{V} \quad (6)$$

SA represents self-attention, Q, K, and V represent Query, Key, and Value respectively. E is the corresponding token dimension, and $B \in \mathbb{R}^{N \times N}$ is the relative position bias, where N is the number of pixels in token. The mathematical expression of Softmax is as follows:

$$\text{Soft max}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (7)$$

Transformer blocks in the encoder and decoder parts of the framework are the same, and both adopt the classical structure of swin-transformer, and the Transformer block in the encoder part distributes Q, K, and V to the Transformer block on the left of the decoder part, as shown in Fig. 1(e). In addition, the encoder part is symmetric, that is, the number of layers of swin-transformer in the same stage number is the same. The same is true of the two parallel swin-transformer in the decoder section. In addition, feature fusion will be performed between the two transformers in the decoder section.

Figure 2 is a schematic of the 3D swin-transformer block clipping window and shifting window. Assuming that the token size is $D \times H \times W$, the window size of partition is $P \times M \times M$, and the token size after partition is $D/P \times H/M \times W/M$. Figure 2(a) and (b) show that 3D square token is cut into several smaller square image blocks by 3D square window, and (b) and (c) are schematic diagrams of interaction of stereo pixel window. In particular, it should be noted that the interaction between windows takes place between windows with a fixed depth range, and there is no information interaction between windows with different depths like Swin UNETR.

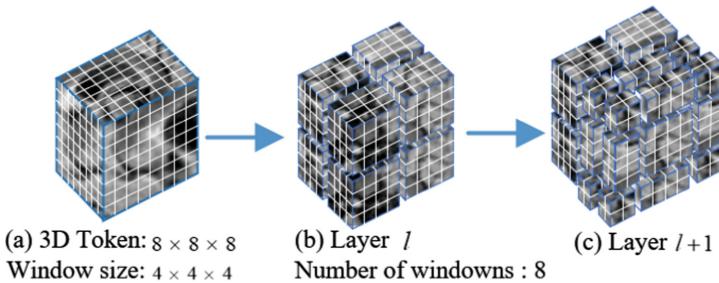


Fig. 2. Shifted windowing mechanism for efficient self-attention computation of 3D tokens with tokens and window size.

After the volume pixel shifting window strategy, the size of token changes from Fig. 2(b) to Fig. 2(c).

3.4 Feature Fusion in Decoder

The fusion of convolutional features and Transformer features is conducted by weighted method after the patch Merging layer, and the weights are learnable. The experiment shows that this fusion method is optimal. In the experiment, we compare the results of

the other two feature fusion methods. Figure 3 is the decoding partial feature fusion structure. Different from the encoder part feature fusion, the decoder part feature fusion adopts linear fusion, and add Fourier Feature Positional Encoding (FPE) for the new fusion Feature.

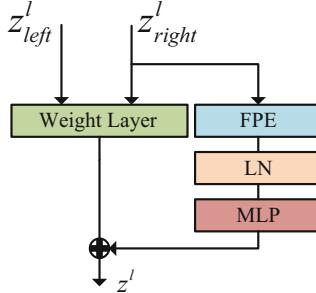


Fig. 3. Fusion Module in decoder. denotes the feature of the left Transformer, and denotes the feature of the right Transformer.

3.5 Classifier Layer

Our goal is to achieve end-to-end segmentation, so the last layer of CST-UNET is a 3D classification layer which map deep E dimensional features to S segmentation classes, and the output tensor size is $D \times H \times W \times S$, Where S is the number of categories of segmentation targets.

3.6 Loss Function

The purpose of model training is to make the model learn the internal relationship between the original data features and the target data features. Finally, when inputting the original data, the model can fit the target data according to the original data, and this method uses the gradient descent method to minimize the loss function continuously. Therefore, the weighted dice coefficient loss function (DL) and cross entropy loss function (CEL) is designed, and the weights are learnable with an initial value of 0.5.

The form of DL is as follows:

$$DL = 1 - \frac{2}{J} \sum_{j=1}^J \frac{\sum_{i=1}^I G_{i,j} Y_{i,j}}{\sum_{i=1}^I G_{i,j}^2 + \sum_{i=1}^I Y_{i,j}^2} \quad (8)$$

The form of CEL is as follows:

$$CEL = -\frac{1}{I} \sum_{i=1}^I \sum_{j=1}^J G_{i,j} \log Y_{i,j} \quad (9)$$

Finally, the Loss used in our task is as follows:

$$\begin{aligned}
 Loss &= \alpha * DL + \beta * CEL \\
 &= \alpha * \left(1 - \frac{2}{J} \sum_{j=1}^J \frac{\sum_{i=1}^I G_{i,j} Y_{i,j}}{\sum_{i=1}^I G_{i,j}^2 + \sum_{i=1}^I Y_{i,j}^2} \right) - \beta * \frac{1}{I} \sum_{i=1}^I \sum_{j=1}^J G_{i,j} \log Y_{i,j}
 \end{aligned} \tag{10}$$

where I denotes the number of pixels, J is the number of categories, $Y_{i,j}$ and $G_{i,j}$ respectively represent the predicted category of pixels and the real category of pixels, α and β are two learnable weight parameters.

4 Experiments

4.1 Dataset and Evaluation Metric

The Brain Tumor Segmentation (BraTS) 2021 [1–3] dataset used in the experiment is a 3D MRI dataset, which is provided by the RSNA-ASNR-MICCAI BraTS 2021 challenge. According to [41], there are 1251 patient samples in this dataset, of which 834 samples are randomly divided for training, 208 samples are used as a validation dataset, and the remaining 209 samples as independent test dataset. Each sample is composed of four modalities of brain MRI scans, namely native T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2) and Fluid Attenuated Inversion Recovery (FLAIR). The pixel of each modality has a uniform volume of $155 \times 240 \times 240$. The labels contain 4 classes: background (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2) and GD-enhancing tumor (label 4). The Dice Score and the Hausdorff Distance (95%) are used to measure the segmentation accuracy of enhancing tumor region (ET, label 1), regions of the tumor core (TC, labels 1 and 4), and the whole tumor region (WT, labels 1, 2 and 4). To eliminate deviations between pixels, a min-max strategy is used before training, and crop the volumes to a fixed size of $128 \times 128 \times 128$ by removing unnecessary background.

4.2 Implementation Details

Python 3.9, PyTorch 1.8.1 and Ubuntu 16.04 are used in this project, with four Tesla V100 GPUs (each has 32 GB memory). The weights of Swin-T pre-trained on ImageNet-1K are used to initialize the model, as well as utilize I3D [42] method inflating 2D swin-transformer and convnets to 3D. For training, we employ Adam optimizer with a learning rate of $1e-4$ for 300 epochs using a cosine decay learning rate scheduler and a batch size of 2. To standardize all volumes, we perform min-max scaling followed by clipping intensity values, and cropping the volumes to a fixed size of $128 \times 128 \times 128$ by excluding unnecessary background (Table 1).

Table 1. The number of parameters and Flops for different frameworks

Method	#param.	Flops
3D U-Net [29]	11.9M	557.9G
V-Net [30]	69.3M	765.9G
VT-UNet [19]	20.8M	134.46G
nnFormer [18]	39.7M	110.7G
TransBTS [4]	33M	333G
UNETR [17]	102.5M	193.5G
Ours	26.98M	154.64G

Our framework has 26.98M parameters and 154.64G Flops. 3D U-NET and V-NET are volume pixel segmentation methods based on convolutional neural network design, VT-UNET and nnFormer are completely based on Transformer architecture, TransBTS and UNETR are deep learning methods based on Transformer and convolution.

Table 2. Our method compares the segmentation results of previous on the BraTS 2021

Method	Dice score ↑				Hausdorff distance ↓			
	ET	TC	WT	AVG	ET	TC	WT	AVG
3D U-Net	83.39	86.28	89.59	86.42	6.15	6.18	11.49	7.94
V-Net	81.04	84.71	90.32	85.36	7.53	7.48	17.20	10.73
TransBTS	80.35	85.35	89.25	84.99	7.82	8.68	5.57	7.35
UNETR	79.78	83.66	90.10	84.51	7.83	8.21	15.12	10.41
nnFormer	82.83	86.48	90.37	86.56	9.72	10.01	15.99	11.90
VT-UNet	85.59	87.41	91.20	88.07	6.23	6.29	10.03	7.52
Ours	85.46	89.38	92.35	89.02	7.95	5.06	4.07	5.69

In the test dataset with 209 samples, the Dice scores of CST-UNet on ET, TC and WT are 85.46%, 89.38% and 92.35% respectively, which are equal to or higher than those of the previous 3D method shown in Table 2.

4.3 Ablation Study and Analysis

Experiments show that setting a reasonable weight for *DL* and *CEL* is of great help to improve the segmentation accuracy of the model, so *DL* and *CEL* weights are set to learnable (Tables 3 and 4).

Table 3. Segmentation results of different weights on the BraTS 2021

DL	CEL	Dice score↑				Hausdorff distance ↓			
		ET	TC	WT	AVG	ET	TC	WT	AVG
0.7	0.3	83.10	87.92	92.19	88.43	15.07	6.96	3.44	8.49
0.4	0.6	82.43	86.99	91.28	87.40	15.38	8.90	7.97	10.75
0.5	0.5	83.81	88.19	92.14	88.60	13.28	7.03	3.88	8.05

↑ means higher is better.

Table 4. Different feature fusion mode and corresponding transmission mode

fusion	way	Dice Score↑				Hausdorff Distance ↓			
		ET	TC	WT	AVG	ET	TC	WT	AVG
L	✓	81.87	83.02	89.86	84.64	10.84	8.89	8.24	9.32
L	✗	82.25	84.65	90.27	85.81	14.26	10.22	7.92	10.80
C	✓	83.41	87.92	91.48	87.96	13.37	7.35	5.67	8.80
C	✗	82.02	83.81	89.82	85.19	14.01	12.37	7.68	11.35

We have explored the fusion of other convolution features and Transformer features and the corresponding transmission paths. *L* represents linear fusion, *C* denotes fusion using convolution layer. ✓ means that the fused features are only fed into the Transformer block at the next layer, the feature fed into CNN at the next layer come from the features of the convolution output at the previous layer, ✗ represents the fused features fed into two branches.

5 Conclusion

A novel U-shaped volumetric medical image segmentation framework based on the preceding ResNet and Transformer framework is proposed. Convolution and Transformer features are fused in the encoder part of the architecture can effectively inject inductive bias acquired by dilated convolution into Transformer architecture, realizing the complementarity of convolution and Transformer architecture. At the same time, the over-fitting risk of Transformer on a small dataset is reduced. In addition, the Transformer architecture in the model is initialized with pre-trained Swin-T parameters on imangenet-1K dataset before training, which effectively improves the segmentation accuracy of the architecture. Additionally, it can directly segment 3D medical images with higher accuracy than previous architectures. In the future, we will continue to study 3D medical image segmentation methods to achieve better segmentation results.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 62172004, 62072002, and 61872004), Educational Commission of Anhui Province (No. KJ2019ZD05).

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data.* **4**, 170117 (2017). <https://doi.org/10.1038/sdata.2017.117>
2. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2015). <https://doi.org/10.1109/TMI.2014.2377694>
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification (2021). <http://arxiv.org/abs/2107.02314>
4. Wang, W., Chen, C., Ding, M., Li, J., Yu, H., Zha, S.: TransBTS: Multimodal Brain Tumor Segmentation using Transformer. [arXiv:2103.04430](https://arxiv.org/abs/2103.04430) [cs] (2021)
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015. LNCS*, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. [arXiv:1706.05587](https://arxiv.org/abs/1706.05587) [cs] (2017)
7. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. *arXiv* (2015). <https://doi.org/10.48550/arXiv.1411.4038>
8. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv* (2016). <https://doi.org/10.48550/arXiv.1511.00561>
9. Wu, Y., He, K.: Group Normalization. [arXiv:1803.08494](https://arxiv.org/abs/1803.08494) [cs] (2018)
10. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) [cs] (2015)
11. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. [arXiv:1702.03118](https://arxiv.org/abs/1702.03118) [cs] (2017)
12. 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. IEEE, Las Vegas, NV, USA (2016). <https://doi.org/10.1109/CVPR.2016.90>
13. Liu, Z., et al.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. [arXiv:2103.14030](https://arxiv.org/abs/2103.14030) [cs] (2021)
14. Dosovitskiy, A., et al.: An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) [cs] (2021)
15. Peng, Z., et al.: Conformer: Local Features Coupling Global Representations for Visual Recognition. [arXiv:2105.03889](https://arxiv.org/abs/2105.03889) [cs] (2021)
16. Chen, J., et al.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. [arXiv:2102.04306](https://arxiv.org/abs/2102.04306) [cs] (2021)
17. Hatamizadeh, A., et al.: UNETR: Transformers for 3D Medical Image Segmentation. [arXiv:2103.10504](https://arxiv.org/abs/2103.10504) [cs, eess] (2021)
18. Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: Interleaved Transformer for Volumetric Segmentation. [arXiv:2109.03201](https://arxiv.org/abs/2109.03201) [cs] (2022)
19. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A Volumetric Transformer for Accurate 3D Tumor Segmentation. [arXiv:2111.13300](https://arxiv.org/abs/2111.13300) [cs, eess] (2021)
20. Wang, Z., Zhang, J., Zhang, X., Chen, P., Wang, B.: Transformer model for functional near-infrared spectroscopy classification. *IEEE J. Biomed. Health Inform.* **1** (2022). <https://doi.org/10.1109/JBHI.2022.3140531>

21. Statistical analysis of multiple significance test methods for differential proteomics. <https://doi.org/10.1186/1471-2105-11-S4-P30>. Accessed 15 May 2022
22. Cheng, M.-T., Ma, X.-S., Zhang, J.-Y., Wang, B.: Single photon transport in two waveguides chirally coupled by a quantum emitter. *Opt. Express, OE.* **24**, 19988–19993 (2016). <https://doi.org/10.1364/OE.24.019988>
23. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized Cut Loss for Weakly-supervised CNN Segmentation. <http://arxiv.org/abs/1804.01346> (2018)
24. Azad, R., Fayjie, A.R., Kauffman, C., Ayed, I.B., Pedersoli, M., Dolz, J.: On the Texture Bias for Few-Shot CNN Segmentation (2020). <http://arxiv.org/abs/2003.04052>
25. Huo, Y., et al.: Fully automatic liver attenuation estimation combining CNN segmentation and morphological operations. *Med. Phys.* **46**, 3508–3519 (2019). <https://doi.org/10.1002/mp.13675>
26. Huang, H., et al.: 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. IEEE, Barcelona, Spain (2020). <https://doi.org/10.1109/ICASSP40776.2020.9053405>
27. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. <http://arxiv.org/abs/1912.05074> (2020)
28. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: D-UNet: a dimension-fusion U shape network for chronic stroke lesion segmentation. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **18**, 940–950 (2021). <https://doi.org/10.1109/TCBB.2019.2939522>
29. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. [arXiv:1606.06650 \[cs\]](arXiv:1606.06650) (2016)
30. 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 (2016). <https://doi.org/10.1109/3DV.2016.79>
31. Vaswani, A., et al.: Attention Is All You Need. [arXiv:1706.03762 \[cs\]](arXiv:1706.03762) (2017)
32. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. [arXiv:2010.04159 \[cs\]](arXiv:2010.04159) (2021)
33. Liu, Z., et al.: Video Swin Transformer. [arXiv:2106.13230 \[cs\]](arXiv:2106.13230) (2021)
34. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A Video Vision Transformer. [arXiv:2103.15691 \[cs\]](arXiv:2103.15691) (2021)
35. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. [arXiv:2102.10662 \[cs\]](arXiv:2102.10662) (2021)
36. Shen, H., Zhang, Y., Zheng, C., Wang, B., Chen, P.: A cascade graph convolutional network for predicting protein-ligand binding affinity. *Int. J. Mol. Sci.* **22**, 4023 (2021). <https://doi.org/10.3390/ijms22084023>
37. Hu, Q., Zhang, J., Chen, P., Wang, B.: Compound identification via deep classification model for electron-ionization mass spectrometry. *Int. J. Mass Spectrom.* **463**, 116540 (2021). <https://doi.org/10.1016/j.ijms.2021.116540>
38. Li, J., Su, Z., Geng, J., Yin, Y.: Real-time detection of steel strip surface defects based on improved YOLO detection network. *IFAC-PapersOnLine* **51**, 76–81 (2018). <https://doi.org/10.1016/j.ifacol.2018.09.412>
39. Xu, Y., Zhang, Q., Zhang, J., Tao, D.: ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias, vol. 14 (2021)
40. Tang, Y., et al.: Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. [arXiv:2111.14791 \[cs\]](arXiv:2111.14791) (2022)

41. Sundaresan, V., Griffanti, L., Jenkinson, M.: Brain tumour segmentation using a triplanar ensemble of U-Nets on MR images. In: Crimi, A., Bakas, S. (eds.) BrainLes 2020. LNCS, vol. 12658, pp. 340–353. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72084-1_31
42. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. IEEE, Honolulu, HI (2017). <https://doi.org/10.1109/CVPR.2017.502>



COVID-19 Classification from Chest X-rays Based on Attention and Knowledge Distillation

Jiaxing Lv¹ , Fazhan Zhu¹ , Kun Lu^{1,2}, Wenyan Wang^{1,2,3} , Jun Zhang⁴ , Peng Chen⁵ , Yuan Zhao^{1,2} , and Ziheng Wu^{1,2()}

¹ School of Electrical and Information Engineering, Anhui University of Technology,
Ma'anshan 243032, Anhui, China
wziheng@ahut.edu.cn

² Ministry of Education, Key Laboratory of Metallurgical Emission Reduction and Resources
Recycling (Anhui University of Technology), Ma'anshan 243002, China

³ School of Materials Science and Engineering, Anhui University of Technology,
Ma'anshan 243032, Anhui, China

⁴ School of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui,
China

⁵ National Engineering Research Center for Agro-Ecological Big Data Analysis and
Application, School of Internet and Institutes of Physical Science and Information Technology,
Anhui University, Hefei 230601, Anhui, China

Abstract. The Coronavirus Disease 2019 (COVID-19) is the pandemic that has had the greatest impact on world economic development in recent years. Early detection is critical to identify patients with COVID-19, chest x-ray is used for early detection is a rapid, extensive and cost-effective method. The existing technology use deep learning methods, and have achieved very good results. However, the training time of deep learning method is long, and the model size makes it difficult to deploy on hardware system. In this work, we have proposed an attention-based ResNet50v2 network, and taken the network as the teacher network to transfer the knowledge to the student network by knowledge distillation. Thus, the student network has higher accuracy and sensitivity to the positive samples of COVID-19 under the condition of low model parameters, high training speed. The experimental results show that our network of teacher and student have achieved 100% accuracy and sensitivity in both COVID-19 and Normal binary classification. In addition, the accuracy rate of teacher network is 98.20%, the sensitivity is 99.58%, the accuracy rate of student network is 97.68%, the sensitivity is 99.17% in the COVID-19, Viral pneumonia and Normal multiple classification, and the parameters of the student network are only 0.269M.

Keywords: COVID-19 · Chest X-ray · Attention-based ResNet50v2 · Knowledge distillation

1 Introduction

The coronavirus Disease (COVID-19) is named as “2019 Coronavirus Disease” by the World Health Organization [1]. Since December 2019, multiple cases of pneumonia of

unknown origin with a history of exposure to the south China seafood market have been identified in some hospitals in Wuhan, Hubei Province, as acute respiratory infections caused by novel coronavirus infection in 2019. On February 11, 2020, WHO director-general Tan Desai has announced in Geneva, Switzerland, the new type of coronavirus infection pneumonia named “COVID-19”, and on March 11th WHO has argued that the current outbreak of COVID-19 could be called a global pandemic [2]. As of Central European Summer Time 5 May 2022 (China Standard Time 6 May 2022), there have been 513384685 confirmed cases of COVID-19 and 6246828 cumulative deaths worldwide [3].

In order to cope with the spread of COVID-19 infection, it is necessary to carry out effective screening and timely medical response to patients. Reverse transcription polymerase chain reaction (RT-PCR) is currently the most common method for clinical screening of patients with COVID-19, which uses respiratory specimens for detection [4]. RT-PCR is used as a reference method for the detection of COVID-19 patients, however, the technique is manual, complicated, laborious and time-consuming [4]. In addition, its supply is prone to shortages, which maybe leads to delay in the disease prevention efforts [5]. In general, COVID-19 infection can be identified by the examination of multifocal and bilateral ground-glass opacity and/or consolidation [6, 7]. Therefore chest x-ray can help early detection of suspected cases [8, 9]. However, radiologists sometimes misdiagnose when hospitals are overworked. Computer-assisted diagnosis (CAD) can diagnose CXR images more quickly and accurately [10–15].

In this work, we hope the model can focus on the important features of the sample. Thus, we built the attention-based ResNet50V2 model to make the model pay more attention to the positive features of COVID-19 and ignore other unimportant features, so as to achieve higher accuracy. And in order to be deployed on hardware system, we have taken the network model as the teacher network to transfer the prior knowledge to the student model, it can improve the accuracy and sensitivity of the student model to COVID-19. The main contributions of this work are summarized as follows:

- An attention-based ResNet50v2 model has been built, the COVID radiology dataset (COVIDRD) was tested for binary classification and multiple classification.
- Student model with attention has the performance of low parameters and fast training speed.
- Teacher model and student model have achieved state-of-the-art recognition accuracy on the COVID radiography dataset (COVIDRD) of the binary classification results.

2 Related Work

The related work of COVID-19 classification models about binary classification, Narin et al. [16] have employed pre-trained ResNet50 model for the three binary classification tasks including normal and COVID-19, normal and viral pneumonia, normal and bacterial pneumonia. Jaiswal et al. [17] have proposed COVIDPEN which is a pruned EfficientNet-based model for COVID-19 classification. Minaee et al. [18] have presented Deep-COVID which is based on deep transfer learning for prediction of COVID-19. Heidari et al. [19] have performed histogram equalization and bilateral low-pass

filter as pre-processing. Then, the classification results have obtained using a transfer learning-based convolutional neural network model. Hemdan et al. [20] have proposed a COVIDX-Net using modified VGG19 model for COVID-19 classification. Afshar et al. [21] have implemented a framework known as COVID-CAPS which is based on a capsule network for COVID-19 classification.

The related work of COVID-19 classification models about multiple classification, the approach of transfer learning in deep learning is utilized by Chowdhury et al. [22] to differentiate between COVID-19 and viral pneumonia based on a dataset acquired from a public database, the models were trained through 423 COVID-19, 1458 viral pneumonia, and 1579 normal chest x-ray images based on augmentation and without augmentation. Mahmud et al. [23] have utilized a deep CNN as COVXNet with modifications based on varying dilation rates for feature extraction, optimization, stacking algorithms, and gradient-based discriminative localization to classify COVID-19 and other types of pneumonia.

3 Method

3.1 Attention-Based ResNet50V2 Model

Our proposed method consists of two constituents which are attention-based ResNet50v2 model and knowledge distillation. Our model consists of three constituents which are ResNet50V2 [24], Convolutional Block Attention Module (CBAM) and classification layer [25], as shown in Fig. 1.

Deep convolutional neural networks work well on image tasks, the superiority of these networks comes from the robust and valuable semantic features they generate from the input images [32, 33]. ResNet50V2 is a modified version of ResNet50 that performs better than ResNet50 and ResNet101 on the ImageNet dataset. In ResNet50V2, a modification is made in the propagation formulation of the connections between blocks. ResNet50V2 also achieves a good result on the ImageNet dataset [26]. In this work, ResNet50V2 as a feature extraction block, the feature extraction block accepts input image size of $224 \times 224 \times 3$ and outputs feature map size of $7 \times 7 \times 2048$.

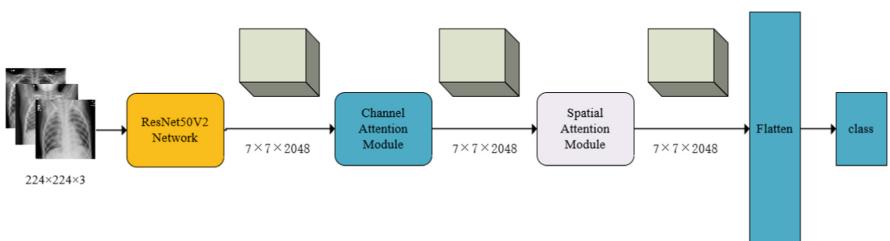


Fig. 1. Attention-based ResNet50v2 network.

CBAM injects attention maps along the channels of the feature map with two independent dimensions of space to increase the representational ability of the network,

focus on important features, and suppress unnecessary features. CBAM can be regarded as Channel Attention Module and Spatial Attention Module in series. Channel Attention Module as shown in Fig. 2, and the computation process can be described as follows.

$$\begin{aligned} M_C(F) &= \sigma(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))) \\ &= \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{max}^c))) \end{aligned} \quad (1)$$

where σ denotes the sigmoid function, $W_0 \in R^{C/r \times C}$, and $W_1 \in R^{C/r \times C}$. Note that the MLP weights, W_0 and W_1 , are shared for both inputs and the Relu activation is followed by W_0 .

The feature map ($7 \times 7 \times 2048$) extracted by ResNet50V2 module is used as input feature F, the input feature F passes through space-based global max pooling and global average pooling for outputting two feature maps ($1 \times 1 \times 2048$), respectively. Then send them into the MLP (two layer neural network), respectively, the number of neurons in the first layer is C/r (C is the channel, r is the reduction rate), the activation is Relu and the network weights of the two layer neural network are shared. Later, the two feature maps output by MLP make the element-wise addition operations, and sigmoid is activated to generate the channel attention feature M_C . Finally, M_C and the input feature map F make the element-wise multiplication operations to generate the input feature F' ($7 \times 7 \times 2048$) required by the Spatial Attention Module.

Spatial Attention Module as shown in Fig. 3, and the computation process can be described as follows.

$$\begin{aligned} M_S(F) &= \sigma(f^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \\ &= \sigma(f^{7 \times 7}([F_{avg}^s; F_{max}^s])) \end{aligned} \quad (2)$$

where σ denotes the sigmoid function and $f^{7 \times 7}$, represents a convolution operation with the filter size of 7×7 .

Taking the feature map F' of the output of the Channel Attention Module as the input feature diagram of this module. First, the feature map F' pass through channel-based global max pooling and global average pooling for outputting two feature maps ($7 \times 7 \times 1$), then the two feature maps make the element-wise concatenation operations. Later, the filter size of 7×7 is made a convolution operation. The dimension is reduced to 1 channel, and sigmoid is activated to generate the spatial attention feature M_S . Finally, the feature M_S and the input feature F' of the module are multiplied to obtain the final generated features F'' .

3.2 Knowledge Distillation

Knowledge Distillation (kd) refers to transferring the prior knowledge from a cumbersome model to a small model [27]. The teacher network is typically much larger and stronger than the student. Under the guidance of teacher network, student network with few parameters, comprehensive training and high recognition rate can be obtained.

Knowledge distillation process can be divided into two steps, the first step is to train teacher network, the second step is to distill the knowledge of teacher network to student network at Temperature = T. The Net-Teacher generated softmax with temperature was

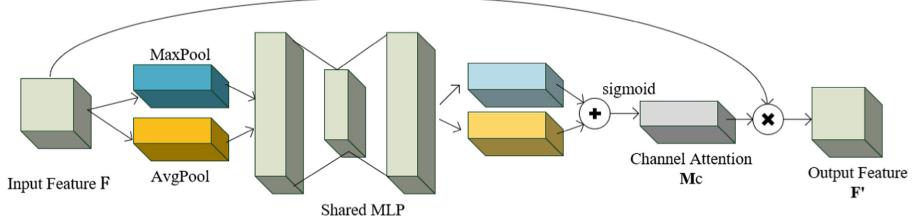


Fig. 2. Structure channel attention module.

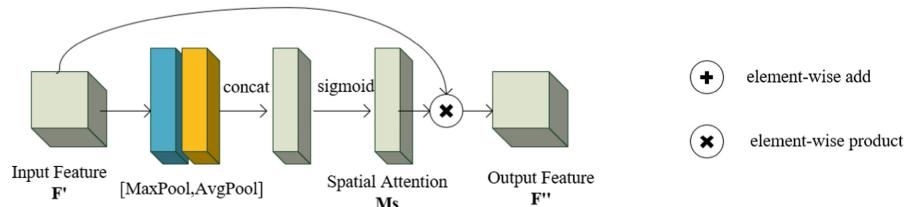


Fig. 3. Structure of spatial attention module.

used as the soft target, Net-Student softmax output at the same temperature and cross entropy of the soft target are the first part loss (L_{soft}) of total loss, the computation process can be described as follows.

$$q^T = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (3)$$

where q^T is the softmax function after adding the temperature variable, z_i indicates the logits for the i -th class, T is a temperature factor.

$$L_{soft} = - \sum_j^N q_{t_j}^T \log(q_{s_j}^T) \quad (4)$$

where $q_{t_j}^T$ is the value of the softmax output at temperature $= T$ of Net-Teacher on class j , $q_{s_j}^T$ is the value of the softmax output at temperature $= T$ of Net-Student on class j .

Net-Student softmax output at the $T = 1$ and cross entropy of the hard target from input images are the second part loss (L_{hard}) of total loss, the computation process can be described as follows.

$$q_s^{T=1} = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (5)$$

$$L_{hard} = - \sum_j^N c_j \log(q_{s_j}^{T=1}) \quad (6)$$

$$L_{total} = \alpha L_{soft} + (1 - \alpha) L_{hard} \quad (7)$$

where c_j is the ground truth value on class j , $c_j \in \{0, 1\}$, select 1 for positive label and 0 for negative label, α is the weight of L_{soft} , $(1 - \alpha)$ is the weight of L_{hard} .

In this work, to obtain a small model with a high sensitivity for COVID-19, we have taken the attention-based ResNet50v2 model as teacher network which has a higher sensitivity for COVID-19, and a nine layers convolutional neural network as student network, achieved knowledge transfer from teacher network to student network, the process as shown in Fig. 4.

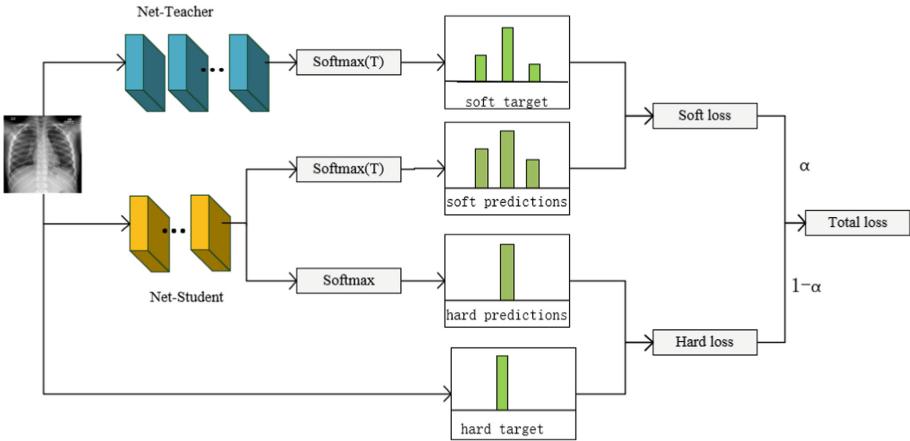


Fig. 4. Knowledge distillation training process.

4 Experiment

4.1 COVID-19 Database and Metric

To evaluate the effectiveness and feasibility of our proposed method, a publicly available database, COVIDRD [31] is used for binary classification and multiple classification in this work.

A team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors have created a database of chest X-ray images for COVID-19 positive cases along with Normal and Viral Pneumonia images. The database contains 1200 images of new coronary pneumonia, 1341 normal and 1345 viral pneumonia Chest x-ray (CXR) images [22, 24], as shown in Fig. 5.

In this work, to make a fair comparison, we randomly select 70% images as the training set, 10% images as the validation set and the remaining 20% images as the test set. To ensure validity of the classifier, we guarantee that the patients used to build the training set and the validation set will not be used in the test set. The evaluation approaches used in the classification results reported in this work are accuracy, sensitivity and specificity, sensitivity and specificity are two proper metrics which can be used for reporting the COVID-19 classified model performance. In this experiment, sensitivity

and specificity represent only metrics of the COVID-19 class, the higher the sensitivity, the stronger the model's ability to discriminate COVID-19, which is defined as follows.

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (9)$$

where TP (True Positive) is the number of correctly classified images of a class, FP (False Positive) is the number of the wrong classified images of a class, FN (False Negative) is the number of images of a class that have been detected as another class, and TN (True Negative) is the number of images that do not belong to a class and did not be classified as that class.

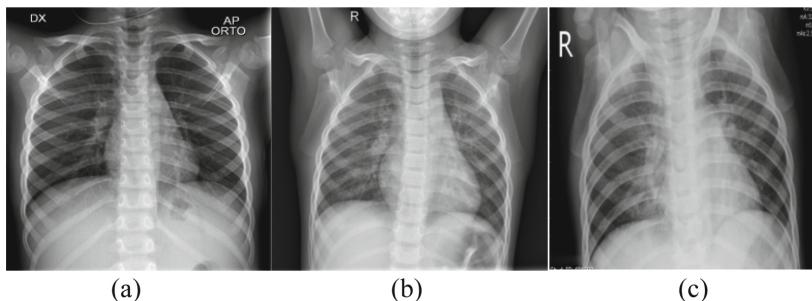


Fig. 5. Chest x-ray for COVID-19 detection, (a) is COVID-19, (b) is Normal, (c) is Viral Pneumonia.

4.2 Experiment Setup

We use the pre-trained model on ImageNet to speed up the training of the model, and take the learning rate decay strategy of cosine annealing to help the model converge to the optimal solution faster. Our experiments are performed using a machine with a Tesla V100 GPU, CUDA 10.2, and cuDNN v9.

4.3 Experiment Result

We extract the feature maps of the output of the feature extraction layer and the attention layer separately to illustrate the localization of features by the attention and no attention models. Grad-CAM is used as a visualization tool to observe our model, as shown in Fig. 6, after adding attention, the model can more precisely locate the location of the decisive feature. Key features can then easily be identified based on where the activation maps are overlapping. And the model performance of the attention models and no attention models are also compared. The result of binary classification as shown in Table 1 and Table 2, the recognition performance of the attention-based models outperform the original models.

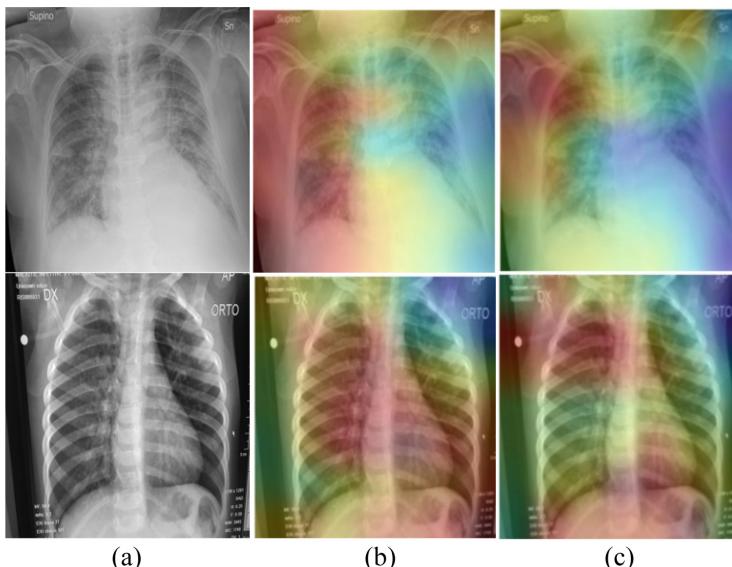


Fig. 6. Positioning of the COVID-19 feature, (a) is original image, (b) is Grad-CAM results at ResNet50V2 module, (c) is Grad-CAM results at Convolutional Block Attention Module.

Table 1. Comparison with attention model and no attention model (binary classification).

Method	Accuracy	Sensitivity	Specificity
ResNet50V2	99.80	99.58	100
Attention-ResNet50V2	100	100	100

Table 2. Comparison with attention model and no attention model (multiple classification).

Method	Accuracy	Sensitivity	Specificity
ResNet50V2	97.55	97.92	100
Attention-ResNet50V2	98.20	99.58	100

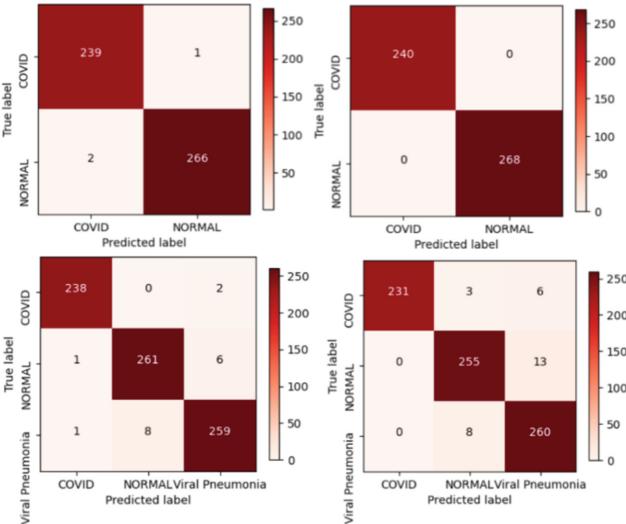
We used the Attention-ResNet50V2 model which is trained consummately on COVIDRD as the Net-Teacher, and the Net-Teacher weights remain constant when performing knowledge distillation. The Net-Teacher's knowledge is transferred to the Net-Student, under the knowledge, Net-Student is trained in the soft target of Net-Teacher and the hard target of ground truth to improve the model performance, as shown in Table 3 and Table 4, after knowledge distillation, the accuracy, sensitivity, and specificity of the Net-Student are all improved in two tasks. And the confusion matrix of binary classification and multiple classification as shown in Fig. 7.

Table 3. Comparison with kd Net-Student and without kd Net-Student (binary classification).

Method	Accuracy	Sensitivity	Specificity
Net-Student (without kd)	99.41	99.58	99.25
Net-Student (kd)	100	100	100

Table 4. Comparison with kd Net-Student and without kd Net-Student (multiple classification).

Method	Accuracy	Sensitivity	Specificity
Net-Student (without kd)	96.13	96.25	96.08
Net-Student (kd)	97.68	99.17	97.01

**Fig. 7.** Confusion matrix of binary classification and multiple classification for Net-Student, the left are the result without kd, the right are the result with kd.

4.4 Comparison with State-of-the-art Methods

Table 5 lists out the performance comparison with the state-of-the-art method in binary classification of COVIDRD. The sensitivity and specificity of the MNRSC model were 99.76% and 99.96%, respectively. The proposed Net-Teacher and Net-Student outperforms the MNRSC model with 100% of sensitivity and specificity. And the number of Net-Student parameters is much smaller than that in other methods. Only 8% of the MNRSC model parameters, as shown in Table 6.

Table 5. Comparison with state-of-art model (binary classification).

Method	Sensitivity	Specificity
TL [18]	98.29	98.02
DL [29]	96.16	97.49
MNRSC [30]	99.76	99.96
Net-Teacher	100	100
Net-Student	100	100

Table 6. Comparison with different model for parameters.

Method	Parameters
GoogleNet	21.905M
InceptionResNet	54.414M
ResNet50	23.788M
MobileNet	3.329M
MNRSC [26]	3.626M
Net-Teacher	28.063M
Net-Student	0.269M

5 Conclusion

The implementation of an effective COVID-19 classification system is still a challenging task due to the recent spreading trend of the COVID-19. This work presents a attention-based model, CBAM is used to enhance the sensitivity of the model to consider attention on important features of COVID-19 and suppress unimportant features. Then, this network model is took as the teacher network to transfer the prior knowledge to the student model, a student network with fewer parameters, faster training, and higher sensitivity is distilled. The COVIDRD datasets has been evaluated to illustrate the effectiveness of our proposed method, the experimental results show that our network of teacher and student have achieved 100% accuracy and sensitivity in binary classification, the parameters of the student network are only 0.269M. And the accuracy rate of teacher network is 98.20%, the sensitivity is 99.58%, the accuracy rate of student network is 97.68%, the sensitivity is 99.17% in multiple classification. The experimental results are important for the design of future computer-aided diagnosis systems. In the future we hope larger datasets from COVID-19 patients become available, make our method more valuable for reference.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (Nos. 62172004, 62072002, and 61872004), Educational Commission of Anhui Province (No. KJ2019ZD05).

References

1. Liu, X., Zhang, S.: COVID-19: face masks and human-to-human transmission. *Influenza Other Respir. Viruses* **14**(4), 472 (2020)
2. Organization, WHO: WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020, Geneva, Switzerland (2020).
3. Organization, WHO: COVID-19 weekly epidemiological update, edn. 84, 22 March 2022 (2022)
4. Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., et al.: Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA* **323**(18), 1843–1844 (2020)
5. Yang, T., Wang, Y.-C., Shen, C.-F., Cheng, C.-M.: Point-of-care RNA-based diagnostic device for COVID-19, vol. 3, p. 165. Multidisciplinary Digital Publishing Institute (2020)
6. Rousan, L.A., Elobeid, E., Karrar, M., Khader, Y.: Chest X-ray findings and temporal lung changes in patients with COVID-19 pneumonia. *BMC Pulm. Med.* **20**(1), 1–9 (2020)
7. Cleverley, J., Piper, J., Jones, M.M.: The role of chest radiography in confirming covid-19 pneumonia. *bmj* **370** (2020)
8. Wang, L., Lin, Z.Q., Wong, A.: COVID-NET: a tailored deep convolutional neural network design for detection of covid-19 cases from chest X-ray images. *Sci. Rep.* **10**(1), 1–12 (2020)
9. Shoeibi, A., Khodatars, M., Alizadehsani, R., Ghassemi, N., Jafari, M., Moridian, P., et al.: Automated detection and forecasting of covid-19 using deep learning techniques: A review. arXiv preprint [arXiv:2007.10785](https://arxiv.org/abs/2007.10785) (2020)
10. Pereira, R.M., Bertolini, D., Teixeira, L.O., Silla, C.N., Jr., Costa, Y.M.: COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Comput. Methods Programs Biomed.* **194**, 105532 (2020)
11. Hammoudi, K., Benhabiles, H., Melkemi, M., Dornaika, F., Arganda-Carreras, I., Collard, D., et al.: Deep learning on chest X-ray images to detect and evaluate pneumonia cases at the era of COVID-19. *J. Med. Syst.* **45**(7), 1–10 (2021). <https://doi.org/10.1007/s10916-021-01745-4>
12. Wang, B., Chen, P., Zhang, J., Zhao, G., Zhang, X.: Inferring protein-protein interactions using a hybrid genetic algorithm/support vector machine method. *Protein Pept. Lett.* **17**(9), 1079–1084 (2010)
13. Wang, B., Valentine, S., Plasencia, M., Raghuraman, S., Zhang, X.: Artificial neural networks for the prediction of peptide drift time in ion mobility mass spectrometry. *BMC Bioinf.* **11**(1), 1–11 (2010). <https://doi.org/10.1186/1471-2105-11-182>
14. Wang, B., Fang, A., Shi, X., Kim, S.H., Zhang, X.: DISCO2: a comprehensive peak alignment algorithm for two-dimensional gas chromatography time-of-flight mass spectrometry. In: Huang, D.-S., Gan, Y., Premaratne, P., Han, K. (eds.) ICIC 2011. LNCS, vol. 6840, pp. 486–491. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-24553-4_64
15. Chen, P., Hu, S., Wang, B., Zhang, J.: A random projection ensemble approach to drug-target interaction prediction. In: Huang, D.-S., Han, K. (eds.) ICIC 2015. LNCS (LNAI), vol. 9227, pp. 693–699. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22053-6_72
16. Narin, A., Kaya, C., Pamuk, Z.: Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal. Appl.* **24**(3), 1207–1220 (2021). <https://doi.org/10.1007/s10044-021-00984-y>
17. Jaiswal, A.K., Tiwari, P., Rathi, V.K., Qian, J., Pandey, H.M., Albuquerque, V.H.C.: COVID-PEN: A novel COVID-19 detection model using chest X-rays and CT scans. *Medrxiv* (2020)
18. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., Soufi, G.J.: Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. *Med. Image Anal.* **65**, 101794 (2020)

19. Heidari, M., Mirniaharikandehei, S., Khuzani, A.Z., Danala, G., Qiu, Y., Zheng, B.: Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int. J. Med. Informatics* **144**, 104284 (2020)
20. Hemdan, E.E.-D., Shouman, M.A., Karar, M.E.: COVIDX-Net: A framework of deep learning classifiers to diagnose covid-19 in X-ray images. *ArXiv preprint arXiv:2003.11055* (2020)
21. Afshar, P., Heidarian, S., Naderkhani, F., Oikonomou, A., Plataniotis, K.N., Mohammadi, A.: COVID-CAPS: a capsule network-based framework for identification of COVID-19 cases from X-ray images. *Pattern Recogn. Lett.* **138**, 638–643 (2020)
22. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., et al.: Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* **8**, 132665–132676 (2020)
23. Mahmud, T., Rahman, M.A., Fattah, S.A.: CovXNet: a multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. *Comput. Biol. Med.* **122**, 103869 (2020)
24. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
25. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018. LNCS*, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012)
27. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv2(7):1503.02531* (2015)
28. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE (2009)
29. Maghdid, H.S., Asaad, A.T., Ghafoor, K.Z., Sadiq, A.S., Mirjalili, S., Khan, M.K.: Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms'. In: *Multimodal Image Exploitation and Learning 2021. International Society for Optics and Photonics*, p. 117340E (2021)
30. Tangudu, V., Kakarla, J., Venkateswarlu, I.B.: COVID-19 detection from chest X-ray using MobileNet and residual separable convolution block. *Soft Comput.* 1–12 (2022)
31. Kaggle COVID-19 radiography database. <https://www.kaggle.com/tawsifurrahman/covid19-radiography-database>. Accessed 18 May 2022
32. Hu, S., Chen, P., Zhang, J., Wang, B.: Prediction of hot spots based on physicochemical features and relative accessible surface area of amino acid sequence. In: Huang, D.-S., Bevilacqua, V., Premaratne, P. (eds.) *ICIC 2016. LNCS*, vol. 9771, pp. 422–431. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42291-6_42
33. Wang, B., Du, L., Zhang, J., Chen, P.: A hierarchical model for identifying mild cognitive impairment. In: *2015 11th International Conference on Natural Computation (ICNC)*, pp. 599–604. IEEE (2015)



Using Deep Learning to Predict Transcription Factor Binding Sites Based on Multiple-omics Data

Youhong Xu^{1(✉)}, Changan Yuan^{2,3}, Hongjie Wu⁴, and Xingming Zhao⁵

¹ Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

1933023@tongji.edu.cn

² Guangxi Academy of Science, Nanning 530007, China

³ Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision, Guangxi Academy Sciences, Nanning 530001, China

⁴ School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

⁵ Institute of Science and Technology for Brain Inspired Intelligence (ISTBI), Fudan University, Shanghai 200433, China

Abstract. Transcription factors (TFs) have a great effect on gene transcription process. TFs can boost the formation of complex gene expression regulation system by promoting or inhibiting gene binding to DNA, which is called as TF binding sites (TFBSs). Recent years have seen the rapid development deep learning (DL) method in natural language processing (NLP), computer vision (CV) and these methods outperform than the state-of-the-art method. Many scholars applied these methods to motif discovery, e.g., DeepBind and DenQ. But these methods only use the raw DNA sequence as input data. Instead of improving complex model, massive biological data brought by high-throughput sequencing technology provides a different idea. In this paper, we propose a simple and effective DL-based model, namely DeepCR, integrating multiple-omics data to predict TFBSs. Experiments on 21 motif datasets of GM12878 cell line from in-vitro protein binding microarray data show that multiple-omics data can significantly improve the overall performance. More specifically, the average AUC is improved by 3.89% for histone modifications, and 3.77% for MeDIP-seq respectively, and 6.63% for histone modifications and MeDIP-seq together. And the mean AR is increased by 3.90% for histone modifications, and 4.50% for MeDIP-seq respectively, and 6.00% for histone modifications and MeDIP-seq together.

Keywords: Transcription factor binding sites · Convolutional neural network · Recurrent neural network · Multiple-omics data · MeDIP-seq · Histone modification · Epigenomic data

1 Introduction

DNA binding proteins (DBPs) refer to proteins which bind to a specific DNA sequence on a chromosome, also known as non-histone proteins. It plays a key role in DNA replication, recombination, strand cleavage, transcription and other processes, and is closely

related to a series of changes in chromatin. TFs [Error! Reference source not found], as an important DBPs, interact specifically with the DNA sequence of the regulatory region and activate or inhibit gene transcription, thereby forming a complex genome expression system. In addition, it has also been reported that abnormal TFs regulation can lead to abnormal expression of downstream genes, causing the occurrence and deterioration of many diseases [2, 3]. Regulating the activity or function of TFs is of great significance for treatment of cancer and autoimmune diseases, so TFs are a potential drug target. TFBSSs are DNA fragments binding to TFs, also known as motif. All these studies tell us, identifying TFBSSs is crucial for further understanding of the transcriptional regulation mechanism in gene expression. A better understanding of binding preferences helps master the transcriptional regulation mechanism, and identifying TFBSSs is the curial step in understanding protein-DNA binding preferences [4].

Second-generation sequencing technology was born in 2005, and can sequence hundreds of thousands to millions of DNA molecules in parallel at a time [5], so it is also known as high-throughput sequencing technology, it provides a large amount of in-vitro binding data to help us study in-vitro protein-DNA binding preferences, such as position weight matrix (PWM). The elements in PWM represent a probability distribution over DNA alphabet {A, C, G, and T} for each position in motif sequence. PWM can express the binding preference of protein and DNA. Therefore, many PWM-based TFBSSs recognition methods have been proposed, which directly learn the binding preference from the original DNA sequence [6–8]. These methods have two shortcomings: one is the length of TFBSS is fixed, another is the nucleotides in the binding site are independently contributed to the calculation of the binding preference. Dependencies between nucleotides can be explicitly encoded by k-mers [9–11], and the result shows that using k-mers as encoding rule is better than PBMs. Although the position-specific sequence nucleus exists, it maps the sequence to higher dimensional space, making these methods inefficient.

In recent years, deep learning offers a scalable, flexible and unified computational approach for pattern discovery. Many new computational methods such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown their superior ability in predicting protein-DNA binding sites [12–18]. These models firstly learn the features by CNNs or RNNs, then use fully connected network to classify TFBSSs. But these models only use raw DNA sequences as input data, many studies have shown epigenomic data [19] may be a nice data supplement to raw DNA sequences. In other words, integrating epigenomic data as input data might help us study in-vitro protein-DNA binding preferences.

In this paper, we first focus on in-depth exploitation of deep CNN and RNN with application to predict TFBSSs in Sect. 2. We call our model DeepCR, which uses CNNs and RNNs extract features from input data, i.e., raw DNA sequences and epigenomic data, and then predicts TFBSSs using fully connected layer. Then we will show experiment results in Sect. 3 and discuss the effect of epigenomic data. At last, we summarize and discuss future work in Sect. 4. The source code and TFs dataset description are available at <https://github.com/suifengwangshi/MotifC>.

2 Materials and Methods

In this section, we first introduce TFs datasets of GM12878 cell line, epigenomic data and data preprocess. Second, architecture of our DeepCR is presented in detail. Third, we give evaluation metrics for DeepCR and hyper-parameters setting.

2.1 Dataset and Preprocessing

We downloaded 21 random TFs datasets (e.g., ATF3, CREB1, etc.,) of GM12878 cell line from the DREAM5 project [20], which comes from a variety of protein families. Each TF dataset comprises a complete set of PBM probe intensities from HK and ME.

2.1.1 DNA Sequence

The development of the HGP (Human Genome Project) and high-throughput sequencing technology have brought us a huge amount of DNA data information. UCSC (The University of California, Santa Cruz) has compiled these data and we can download these data from the website (<https://hgdownload.soe.ucsc.edu/downloads.html>).

With the development of technology, the human genome data has also developed from hg4 in 2000 to hg38 in 2013. The hg19 data in 2009 is used in this article. The size is 3 GB, and each base letter (A, G, C, T, N, the first four correspond to the four nucleotides that make up DNA, and N represents the unknown bases restricted by sequencing technology and experiments) corresponds to one byte, so it is 3 billion bp (base pair), including 22 autosomes and X, Y sex chromosomes and M mitochondrial chromosomes.

2.1.2 Epigenomic Data

Transcription factors play an important role in gene expression and can guide gene transcription and protein synthesis. In this way, genome data in organisms can also be used as guidance data to improve the model's recognition of transcription factors, Methylated DNA immunoprecipitation sequences (MDS) and histone modifications (HMS) are selected from the ENCODE Epigenetics dataset.

MeDIP-Seq sequencing is a genome-wide methylation detection technology based on antibody enrichment principle. MeDIP technology is used to specifically enrich the methylated DNA fragments on the genome through 5'-methylcytosine antibody, and then high-throughput sequencing can carry out high-precision CpG intensive high methylation region research at the genome-wide level. Researchers can use medip SEQ technology to quickly and effectively find methylation regions in the genome, so as to compare the differences of DNA methylation modification patterns between different cells, tissues or disease samples.

Histone is the target of epigenetic modification (including methylation and ubiquitination), which has the function of regulating gene expression. The study of histone modification information can further study gene expression.

2.1.3 Data Encoding

One-hot code, also known as a valid code, uses N-bit status to encode N states. Each state has its own independent bit. And at any time, only one is valid. We use one-hot code to express DNA sequences as following Table 1.

Table 1. DNA sequence one-hot corresponding code

Nucleotide	Encoder
A (adenine)	(1, 0, 0, 0)
C (cytosine)	(0, 1, 0, 0)
G (guanine)	(0, 0, 1, 0)
T (thymine)	(0, 0, 0, 1)
N (unknown)	(0, 0, 0, 0)

Assuming a DNA sequence of length L is represented as $S = (s_1, s_2, \dots, s_L)$, we can get a $L * 4$ matrix through one-hot encoding. In addition, we add MeDIP-seq (MDS) and histone modifications (HMS) of each base of the input sequence. Thus, each input sequence S with n nucleotides is encoded as $n \times 6$, meaning sequence with length L and each has 6 channels. The first 4 channels are for one-hot encoding and the other 2 channels for MDS and HMS respectively.

2.1.4 Dataset Construction and Division

The length of TFBSs is usually 5–20 bp, so the length of the input sequence is selected as 101 bp. According to the ratio of 1:1, the positive sample and the negative sample are selected. The positive sample data is centered on the binding site, and the sequence length is selected as 101 bp. The bases before and after are used to provide contextual information. The negative sample is selected about 3000 bp behind the binding site. The number of positive and negative samples of the 21 TFs selected on the GM12878 cell line is shown in the Table 2 below.

Table 2. Numbers of positive and negative samples.

TFs	Positive sample	Negative sample
ATF3	2921	2921
BATF	52216	52216
BCL11A	31672	31672
CEPB	12247	12247
CREB1	15758	15758
EGR1	17468	17468

(continued)

Table 2. (*continued*)

TFs	Positive sample	Negative sample
ELF1	31332	31332
ETS1	8436	8436
MEF2A	19145	19145
PBX3	34215	34215
POU2F2	45368	45368
RUNX3	90134	90134
SP1	37020	37020
SRF	10420	10420
STAT5A	30266	30266
TAF1	17585	17585
TCF12	40188	40188
USF1	9268	9268
YY1	43592	43592
ZBTB33	13572	13572
ZEB1	17451	17451

Considering numbers of sample is adequate and to accurately evaluate the model's performance, 5-fold cross-validation strategy [21] was adopted in this paper. Dataset is randomly divided into 5 equal parts, and four parts used as the training data while the rest used as the test data. This strategy repeated k times in total, and take the average as the final result.

2.2 Network Architecture

DeepBind [12] is the first to apply deep learning to motif discovery, using a convolutional layer to extra feature from DNA sequence and fully connected layer to classify TFBSs. Comparing to machine learning methods, DeepBind has greatly improved performance and was proved to be a successful attempt.

Considering only using one layer to extract features from complex data may be not enough, we proposed a deeper neural network model combining CNNs and RNNs, namely DeepCR. Convolutional layers are accompanied by activation function, dropout and local pooling strategies. Recurrent layers contain several hidden layers to reduce data dimension. Convolutional layers can compute a score for all potential local motif from input data. Deeper layers are able to learn more fully features comparing to DeepBind. Recurrent layers are followed, in the hope that it can capture the interaction pattern in neighboring sequence. Recurrent layers take the motif score sequence computed by convolution layers as input and recognize the distribution pattern of the motif score sequence. That is to say, recurrent layers followed by convolutional layers can take the

interaction of the local motifs into consideration. Deeper convolutional layers improve the receptive field and recurrent layers have the function of memory for surrounding data. All of these allow an overall pattern recognition of the candidate sequence. Each convolution layer is followed by a max-pooling layer and a dropout layer. Max-pooling layer could simplify features to use only representative ones. Dropout layer can randomly discard some neurons, preventing them from propagating data backwards, to light the overfitting risk. Finally, extracted features are fed into fully connected layers to classify original sequence.

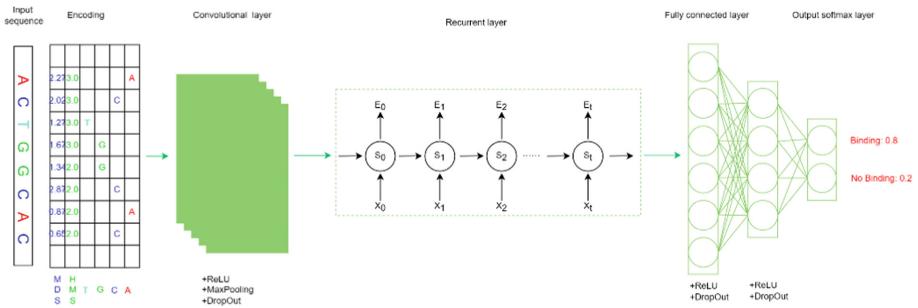


Fig. 1. An overview of the DeepCR model.

Because DNA sequence is one dimension, so the convolution layer in our paper is a one-dimensional convolution expressed in Eq. (1). I mean input data, o is the indices of the output position and k is the k^{th} the kernel, and W^k represents the weight matrix of the k^{th} kernel whose shape is $S \times N$, S filter size and N input channels.

$$X_o^k = \sum_{m=0}^{S-1} \sum_{n=0}^{N-1} I_{o+m,n} * W_{m,n}^k \quad (1)$$

For recurrent layer, hidden layer node is calculated by Eq. (2) as follow. U and W are the weights of input x and output h respectively. f is the activation function and b mean the bias. It can be seen that when calculating h_t , the output state h_{t-1} of previous data is also involved.

$$h_t = f(U * x_t + W * h_{t-1} + b) \quad (2)$$

Also, h_{t-1} can be calculated by Eq. (3) below.

$$h_{t-1} = f(U * x_{t-1} + W * h_{t-2} + b) \quad (3)$$

Combining Eq. (2) and Eq. (3), we can get Eq. (4). It tells us that the t^{th} data have previous memory, so recurrent layer can process sequence data.

$$h_t = f(U * x_t + W * f(U * x_{t-1} + W * h_{t-2} + b) + b) \quad (4)$$

Adding dropout after fully connected layer results in Eq. (5) where m_i is sampled from Bernoulli distribution.

$$z_m = w_{d+1} + \sum_{i=1}^d m_i * w_{i,m} * y_i \quad (5)$$

The rectified linear unit activation function is used in this design and it is given in Eq. (6). ReLU function introduces non-linear features to DeepCR model.

$$\text{ReLU}(x) = \begin{cases} 0, & x < 0 \\ x, & \text{others} \end{cases} = \max(0, x) \quad (6)$$

The final layer is the softmax layer that normalizes its input vector z into a probability distribution having M probabilities proportional to the exponential of the input numbers, expressed by Eq. (7).

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{m=1}^M \exp(z_m)} \quad (7)$$

Figure 1 plots a graphical illustration of DeepCR and the detailed parameter settings in each layer are listed in [Error! Reference source not found]. Output shape is represented as (B, L, N) , B to batch size, L to length and N to channels. Input data is $(B, 101, 6)$. It should be mentioned that hyper-parameter settings inherent from classic deep learning methods and choose from grid search in training procedure.

2.3 Loss Function and Evaluation Metric

Classifying TFBSSs is two-class work and the ratio of positive and negative samples with 1:1, so we use cross binary entropy loss function as followed Eq. (8). y_i represents the label of sample i , and the positive sample is 1 while the negative is 0. p_i is the probability that sample i is predicted to be positive.

$$L = \frac{1}{N} \sum_i L_i = \frac{1}{N} \sum_i -[y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)] \quad (8)$$

In the binary classification, it is generally said that the real category whose label is 1 is positive, while the real category whose label is 0 is negative. If the prediction is correct (wrong), the result is true (false). Combining the above four cases, we can get the confusion matrix shown in Table 4.

Our model uses AUC (Area Under the Curve of Receiver Operating Characteristic curve, ROC-AUC) and Accuracy as metric evaluation. We can calculate Accuracy, TPR (true positive ratio) and FPR (false positive ratio) from Table 4 using formula below. Accuracy rate (AR) refers to the proportion of correctly classified data to the total numbers of data. TPR refers to the proportion of data whose actual category is positive that is predicted to be positive. FPR refers to the proportion of data whose real category is negative that is predicted to be positive.

$$AR = \frac{TP + TN}{TP + FN + FP + TN} \quad (9)$$

Table 3. Parameter setting of DeepCR model in detail.

Architectures	Settings	Output shape
Input data	-----	(B, 101, 6)
1 st conv layer	kernel number = 64, size = 15, stride = 1, padding = 0	(B, 87, 64)
ReLU layer	-----	(B, 87, 64)
Pooling layer	kernel size = 4, stride = 4, padding = 0	(B, 21, 64)
Dropout layer	ratio = 0.2	(B, 21, 64)
2 nd conv layer	kernel number = 64, size = 5, stride = 1, padding = 0	(B, 16, 64)
ReLU layer	-----	(B, 16, 64)
Pooling layer	kernel size = 4, stride = 4, padding = 0	(B, 4, 64)
Dropout layer	ratio = 0.2	(B, 4, 64)
Recurrent layer	in = 256, hidden = 64, out = 64	(B, 64)
1 st fc layer	dim = 64, regularization = ‘L2’	(B, 64)
ReLU layer	-----	(B, 64)
Dropout layer	ratio = 0.2	(B, 64)
2 nd fc layer	dim = 1	(B, 1)
Softmax layer	For binary classification, degenerates to sigmoid layer	(B, 1)

Table 4. Confusion matrix.

		Predicted category	
		True (1)	False (0)
Real category	Positive (1)	True positive sample (TP)	False positive sample (FP)
	Negative (0)	False negative sample (FN)	True negative sample (TN)

$$TPR = \frac{TP}{TP + FN} \quad (10)$$

$$FPR = \frac{FP}{FP + TN} \quad (11)$$

AUC represents the area of the area between ROC curve and the horizontal axis, and its value is between 0 and 1. The specific meaning refers to the probability value of the positive sample predicted is greater than the predicted probability of a negative sample. The larger the AUC value, the better the effect of the model.

2.4 Experiment Setting

Weights are initialized by Xavier uniform initializer, and optimized by Adam algorithm with batch-size of 100. We implement grid search strategy over some sensitive hyper-parameters, i.e., dropout ratio, L2 weight decay, and momentum. Detailed hyper-parameter setting is listed in Table 5.

Table 5. A list of sensitive hyper-parameters and grid search space in experiment.

Hyper-parameters	Settings
Dropout ratio	0.2, 0.5
Learning rate	0.001
Momentum	0.999, 0.99, 0.9
Weight decay	5E-4, 1E-3, 5E-3
Epoch	20
Batch size	100
Threshold	0.9

3 Results and Analysis

3.1 Results Display

In order to verify the effectiveness of MeDIP-seq (MDS) and histone modifications (HMS), we train the DeepCR model using raw DNA sequences only. For a fair comparison, we all use best hyper-parameters by grid search in Table 3. We use different data as model input, i.e., raw DNA sequences, raw DNA sequences + MDS, raw DNA sequences + HMS, raw DNA sequences + MDS + HMS respectively. The result of comparison is illustrated in Fig. 2, and Fig. 3.

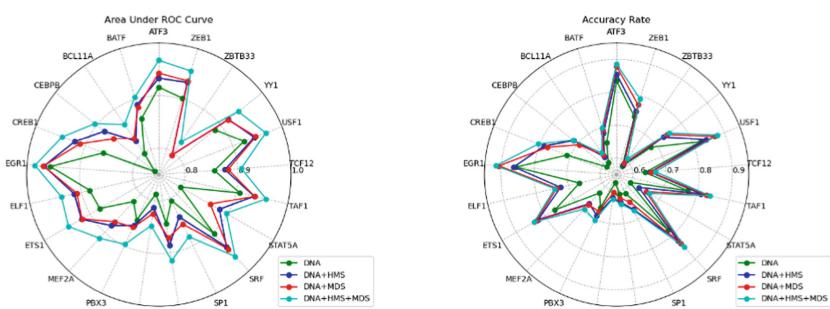


Fig. 2. AUC (left) and AR (right) of DeepCR model plus HMS and MDS

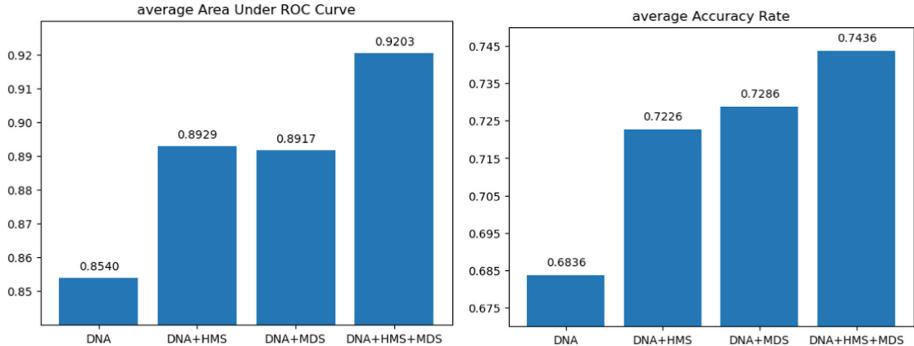


Fig. 3. Average AUC (left) and AR (right) of DeepCR model plus HMS and MDSA

3.2 Effect of HMS and MDS

Figure 2 shows that AUC and AR of randomly selected 21 TFs, and Fig. 3 is average metric. Model adding HMS or MDS encircles without adding data. We can draw two results from above ablation experiment: (1) adding HMS and MDS information as additional input data can greatly improve AUC and AR. Besides, when input data contains HMS and MDS, AUC and AR outperform only containing HMS or MDS. (2) In term of concrete effort, the average AUC of using raw DNA sequence only is 85.40% while it is 89.29% integrating HMS, and 89.17% integrating HMS to raw DNA sequences respectively. On the other hand, the mean AR of using raw DNA sequences only is 68.36% while it is 72.26% integrating HMS, and 72.86% integrating MDS to raw DNA sequences respectively. Thus, adding HMS to the raw DNA sequences improve the performance by 3.89% and 3.90% in terms of AUC and AR respectively, MDS by 3.77% and 4.50%. Then we conduct experiments integrating HMS and MDS to raw DNA sequences, and the average AUC is 92.03% comparing 85.40% and the average AR is 74.36% comparing 68.36%. There is 6.63% increase to average AUC and 6.00% to average AR.

4 Conclusion and Future Work

Motif discovery is an important step for a better studying of molecular and cellular biology. In this paper, we propose a simple and efficient model combining convolutional network and recurrent network, namely DeepCR for predicting TFBSs, integrating multiple-omics data (i.e., HMS and MDS) with raw DNA sequences. Integrating these data to DNA sequences respectively can promote the average AUC and AR, and while including HMS and MDS together to raw DNA sequences, we can get better result comparing only any data.

Although integrating multiple-omics data has some good performance to predict TFBSs, there are some obvious shortcomings in future work: (1) why HMS and MDS data have good performance, and if we train model using any data while test using another, whether good results exist. (2) Fully connected layer limits the size of input data. If we can use the pooling layer instead of the full connected layer, we can solve

the problem of input data size. (3) Different code rules for input data also can influence results [22–26], as we know, encoding input to embedding vector is a commonly used data-preprocessing way for sparse data. (4) DNA shape is an important replenish for predicting TFBSS [27, 28]. (5) Due to the similarities between DNA and RNA, we could some methods predicting RNA-Protein binding preferences to predict TFBSSs [29–33].

Acknowledgements. This work was supported by the grant of National Key R&D Program of China (No. 2018YFA0902600 & 2018AAA0100100) and partly supported by National Natural Science Foundation of China (Grant nos. 61732012, 62002266, 61932008, and 62073231), and Introduction Plan of High-end Foreign Experts (Grant no. G2021033002L) and, respectively, supported by the Key Project of Science and Technology of Guangxi (Grant no. 2021AB20147), Guangxi Natural Science Foundation (Grant nos. 2021JJA170204 & 2021JJA170199) and Guangxi Science and Technology Base and Talents Special Project (Grant nos. 2021AC19354 & 2021AC19394).

References

1. Lambert, S.A., et al.: The human transcription factors. *Cell* **175**(2), 598–599 (2018)
2. Teixeira, J.R., Szeto, R.A., Carvalho, V.M.A., et al.: Transcription factor 4 and its association with psychiatric disorders. *Transl. Psychiatry* **11**(1), 1–12 (2021)
3. Wu, Q., Li, W., You, C.: The regulatory roles and mechanisms of the transcription factor FOXF2 in human diseases. *PeerJ* **9**, e10845 (2021)
4. Tianyin, Z., Ning, S., et al. Quantitative modeling of transcription factor binding specificities using DNA shape. In: Proceedings of the National Academy of Sciences, pp. 112–115 (2015)
5. Schuster, S.C.: Next-generation sequencing transforms today's biology. *Nat. Methods* **5**(1), 16–18 (2008)
6. Storto, G.D., Zhao, Y.: Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.* **11**(11), 751–760 (2010)
7. Bi, Y., Kim, H., Gupta, R., et al.: Tree-based position weight matrix approach to model transcription factor binding site profiles. *PLoS One* **6**(9), e24210 (2011)
8. Giaquinta, E., Grabowski, S., Ukkonen, E.: Fast matching of transcription factor motifs using generalized position weight matrix models. *J. Comput. Biol.* **20**(9), 621–630 (2013)
9. Fletez-Brant, C., Lee, D., McCallion, A.S., et al.: kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* **41**(W1), W544–W556 (2013)
10. Ghandi, M., Lee, D., Mohammad-Noori, M., et al.: Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**(7), e1003711 (2014)
11. Lee, D.: LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**(14), 2196–2198 (2016)
12. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015)
13. Jian, Z., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**(10), 931–934 (2015)
14. Zhang, Q., Zhu, L., Bao, W., Huang, D.-S.: Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**(2), 679–689 (2020)
15. Zhang, Q., Zhu, L., Huang, D.-S.: High-order convolutional neural network architecture for predicting DNA-protein binding sites. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**(4), 1184–1192 (2019)

16. Zhang, Q., Shen, Z., Huang, D.-S.: Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. *Sci Rep.* **9**(1), 8484 (2019)
17. Zhang, H., Zhu, L., Huang, D.S.: DiscMLA: an efficient discriminative motif learning algorithm over high-throughput datasets. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**(6), 1810–1820 (2018)
18. Zhu, L., Zhang, H., Huang, D.S.: LMMO: a large margin approach for optimizing regulatory motifs. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **15**(3), 913–925 (2018)
19. Ritambhara, S., Lanchantin, J., et al.: DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32**, i639–i648 (2016)
20. Weirauch, M.T., Cote, A., Norel, R., et al.: Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**(2), 126–134 (2013)
21. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI* **14**(2), 1137–1145 (1995)
22. Wang, J., Huang, P., Zhao, H., Zhang, Z., Zhao, B., Lee, D.L.: Billion-scale commodity embedding for E-commerce recommendation in Alibaba. In: *Knowledge Discovery and Data Mining*, pp. 839–848 (2018)
23. Zhu, L., Guo, W.-L., Huang, D.-S., Lu, C.-Y.: Imputation of ChIP-seq datasets via low rank convex co-embedding. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 141–144 (2015)
24. Wang, D., Zhang, Q., Yuan, C.-A., Qin, X., Huang, Z.-K., Shang, L.: Motif discovery via convolutional networks with K-mer embedding. In: Huang, D.-S., Jo, K.-H., Huang, Z.-K. (eds.) *ICIC 2019. LNCS*, vol. 11644, pp. 374–382. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26969-2_36
25. Zhu, L., Guo, W.-L., Huang, D.-S., Lu, C.-Y.: Imputation of ChIP-seq datasets via Low Rank Convex Co-Embedding. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 141–144 (2015)
26. Wenzuan, X., Zhu, L., Huang, D.-S.: DCDE: an efficient deep convolutional divergence encoding method for human promoter recognition. *IEEE Trans. Nanobiosci.* **18**(2), 136–145 (2019)
27. Zhang, Q., Shen, Z., Huang, D.-S.: Predicting in-vitro transcription factor binding sites using DNA sequence + shape. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**(2), 667–676 (2021)
28. Wang, S., He, Y., Chen, Z., Zhang, Q.: FCNGRU: locating transcription factor binding sites by combining fully convolutional neural network with gated recurrent unit. *IEEE J. Biomed. Health Inform.* **26**(4), 1883–1890 (2022)
29. Shen, Z., Zhang, Q., Han, K., Huang, D.-S.: A deep learning model for RNA-protein binding preference prediction based on hierarchical LSTM and attention network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**(2), 753–762
30. Shen, Z., Deng, S.-P., Huang, D.-S.: Capsule network for predicting RNA-protein binding preferences using hybrid feature. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**(5), 1483–1492 (2020)
31. Shen, Z., Deng, S.-P., Huang, D.-S.: RNA-protein binding sites prediction via multi scale convolutional gated recurrent unit networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**(5), 1741–1750 (2020)
32. Shen, Z., Bao, W., Huang, D.-S.: Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep.* **8**(1), 15270 (2018)
33. Shen, Z., Zhang, Y.-H., Han, K., Nandi, A.K., Honig, B., Huang, D.-S.: miRNA-disease association prediction with collaborative matrix factorization. *Complexity* **2017**(2017), 1–9 (2017)



Non-invasive Haemoglobin Prediction Using Nail Color Features: An Approach of Dimensionality Reduction

Sunanda Das, Abhishek Kesarwani, Dakshina Ranjan Kisku^(✉), and Mamata Dalui

Department of Computer Science and Engineering, National Institute of Technology Durgapur, Durgapur, West Bengal, India

{sd.19cs1111, ak.18cs1102}@phd.nitdgp.ac.in, {drkisku, mamata.dalui}@cse.nitdgp.ac.in

Abstract. Estimation of blood haemoglobin level is essential for evaluating health condition related to anaemia and its associated diseases. The invasive way of haemoglobin estimation is costly as well as it needs trained professionals along with a good infrastructural and administrative support. Due to the lack of specialized equipment, trained professionals, and laboratory facilities, non-invasive haemoglobin estimation approaches are in high demand, specially, in the rural areas where there is lack of resources. This paper reports a smartphone camera-based non-invasive haemoglobin prediction model that analyses the video photography of pallor changes of the nail bed due to pressure application and release. Unlike different color space features extracted directly from the video, a reduced dimensional latent space representation of those features is considered to train a model for haemoglobin level estimation. A Gaussian-neighborhood function in error calculation of autoencoder is incorporated to get a reduced dimensional latent structure that holds all the properties of input patterns. The mean root-mean-squared (RMSE) error between the actual and predicted output derived from the different state-of-the-art regression models and a weighted sum rule over the output of regression models are compared to prove the efficacy of the proposed model.

Keywords: Anaemia · Haemoglobin · Autoencoder · Dimensionality reduction

1 Introduction

Anaemia is a condition having lesser amount of haemoglobin components in blood. According to the World Health Organization (WHO) statistics, the prevalence of anaemia is considerably high in developing countries and less developed countries. Haemoglobin concentration for healthy men is normally 13.5 g/dl to 17.5 g/dl and that for women is 12 g/dl to 15.5 g/dl, though it varies with age, body weight, pregnancy status, and geographical region. Due to the red color of haemoglobin, a very common age-old practice for examining whether a person is anaemic or not, is by observing the redness in eye conjunctiva, skin, tongue, and fingertip. For accurate estimation, laboratory-based haemoglobin tests are prescribed though they are costly, painful as well as time-consuming. In addition

to that, the invasive haemoglobin estimation approaches need trained professional, laboratory equipment that may not be easily available in the rural areas. Thus, non-invasive approaches are in high demand for predicting the haemoglobin level instantly, specially, where there is a need for regular monitoring of blood haemoglobin concentration for evaluation of anaemia and other related health issues.

Existing non-invasive approaches for estimation of blood haemoglobin level, mainly focus on analyzing the images of eye conjunctiva, tongue, palm pallor, nail pallor or analyzing the Photoplethysmography (PPG) signals extracted from fingertips. Some of the existing methods that deal with the images, choose their region of interest area by manual selection that may lead to inaccurate result due to inappropriate selection of region of interest. However, methods that use PPG signals, are indeed costly due to huge resource requirements. Most of the existing techniques observe and analyse the correlation between the haemoglobin level and extracted features from the image, mostly a single image. This provides an erroneous output if the images are not captured properly, having inconsistent lighting conditions, inappropriate camera focus, etc. Other than those drawbacks, existing approaches apply the features extracted from either images or signals, directly to the regression models. Though the extracted features have great impact over the predicted haemoglobin level; however, it may contain some redundant and non-essential feature values that degrade the overall performance of the model. In view of the current scenario, a non-invasive haemoglobin prediction system based on color channel features extracted from the video photography of the nail bed, wherein the extracted features have high correlation with blood haemoglobin concentration, has been proposed. The extracted features exhibit paramount importance in order to estimate the haemoglobin level. However, some redundant and irrelevant features may appear due to which the system may suffer from degraded performance. To get rid of redundant and multi-collinear features, dimensionality reduction techniques are considered in order to decrease the dimension of feature space into a meaningful feature set. A higher dimensional feature set reveals higher degrees of freedom that may overfit the training data and does not perform well on the test data. The distance between the nearest and farthest data points can also become equidistant in higher dimensions, which affect the accuracy of some distance-based analysis models. On reducing the feature dimension, the model can ensure good generalization properties. It also uses a lesser number of degrees of freedom which can take care of the overfitting problem. Among many other dimensionality reduction approaches, autoencoder is one of the methods that reconstructs the input layer architecture in the output layer. It is capable to produce a lower-dimensional latent space that contains more accurate and necessary information and holds the pattern or structure of the original feature space. The latent space representation provides a non-linear structure, and is capable to capture complex patterns and sudden changes in values. In the proposed methodology, the autoencoder is used for dimensionality reduction by removing the anomalies in data and to yield a non-linear transformation that is more informative.

The manuscript is organized as follows. Section 2 presents a brief overview of the related works. The proposed work methodology has been detailed out in Sect. 3 followed by experimental results in Sect. 4. Last section concludes the work.

2 Related Works

There are a handful of works which have been presented in literature towards measuring the blood haemoglobin level, in non-invasive way by considering some physiological characteristics of human body. For example, Muhe et al. [1] have presented a study where pallor of eye conjunctivae, tongue, palm, and nail are clinically examined, and determined the types of anaemia by comparing with clinical haemoglobin values. In [2], authors estimated the haemoglobin level by analyzing the average R, G, B values collected from the images of blood sample and train a Back Propagation Network for haemoglobin level prediction. Kavsaoglu et al. [3] have discussed a non-invasive method to determine the blood haemoglobin level by exploring the PPG signal taken from the right index finger of the user. They have extracted 44 different features which are considered with the user's height, weight, sex, age etc. and predict the haemoglobin level by different regression models. Authors in [4] have correlated the haemoglobin value with the mean intensity values of RGB channels of the ROI selected region of the images of lower eyelid. The ROI is selected manually which may affect the accuracy if the proper ROI is not chosen. Roy Chowdhury et al. [5] have analyzed pallor sites of eye conjunctiva and tongue to screen anaemia and classified into different classes based on the use of the color plane, intensity, and gradient feature. In [6], a non-invasive anaemia detection method has been reported wherein authors have examined the skin color (redness) of the palm by capturing video of blood flow after occlusion and then after release. After formation of frame, ROI is selected and extracted features are used to train a linear regression model to find out the haemoglobin level. Another non-invasive technique for the detection of anaemia is reported in [7] where pictures of lower anterior eye conjunctiva are captured, and a specific area of interest is chosen. The difference between average red pixel intensity values and average green pixel intensity values is determined and establishes a correlation between the intensity differences with only the class of anaemic or non-anaemic subjects. However, the variation in eye sizes is reflected in the ROI selected region that may create a deviation of the predicted haemoglobin level from the clinical estimation. In [8], authors have presented a non-invasive approach for anaemia detection by analyzing digital photographs of eye conjunctiva using a smartphone-based camera. At first, color calibration is applied to adjust the brightness of images, then conjunctival erythema index (EI) is calculated from R, G, B color scales and a correlation between EI index and clinically tested hemoglobin concentration is delineated. In [9], a smartphone-based application for anaemia detection has been proposed based on digital images of the nail bed. In this work, photos of all fingers are captured, and ROI is selected. Then an average of different features is calculated to train a multi-linear regression model. The accuracy of this approach can decrease due to the non-uniform illumination condition. Hema-App [10], a popular smart-phone based application, following the Beer-Lambert law aims to measure the ratio (IR) between the maximum and minimum intensity values absorbed by the Hemoglobin and plasma. After gathering the video, the high pass filter is used to remove fluctuation due to breathing. Then FFT is applied to the filtered waveform to find out the dominant peak, which provides an estimated value of heart rate. They extract other features like peak index, troughs index, IR values, and their ratios to train the Support Vector Regression (SVR) model to predict haemoglobin level. In [11] anaemia is detected using CNN by analyzing the eye conjunctiva palpebral images. Naik *et al.*

[12] proposed another artificial intelligence-based approach for anaemia detection using ECG and PPG signals. Acharya et al. [13] have invented a non-invasive approach for haemoglobin level prediction where PPG signals from the finger are exploited. However, the method is costly. In another literature [14] PPG signals are also considered for non-invasive anaemia detection by applying the PCA techniques for better selection of features. To overcome the limitations of the invasive systems as well as state-of-the-art non-invasive systems, a low-cost, non-invasive, user-friendly system for accurate estimation of blood haemoglobin level has been proposed here.

The proposed system estimates the haemoglobin level by analyzing the variations of color of the nail bed of hand due to the pressure application and release. The characteristics of color with different time stamps are video photographed with a smartphone camera along with a customized nail device and is analyzed to extract different features related to the different color channels. To enhance the inter-relationship among the features, a latent feature space representation is also presented. This latent space structure increases the efficacy of the system when they are fed into the regression models. A set of regressions models along with a weighted sum approach is applied to generate the final predicted output. The next section details out the proposed methodology. The proposed method is based on the nail pallor as pressure application & release is not possible in case of tongue and eye conjunctiva. And our main aim is to estimate haemoglobin by analyzing nail pallor by pressure application.

3 Proposed Methodology

The proposed methodology is proficient to estimate the haemoglobin level non-invasively by analyzing the color-changing scenario of the nail bed by applying and releasing pressure. Though color of tongue, eye conjunctiva provide good information in case of haemoglobin estimation, but pressure application is not possible in those part of body. The core idea of the proposed method is to deoxygenate the stagnant blood in nail bed tissues by applying pressure, and thereafter to allow quick reoxygenation by suddenly releasing the compression to observe the color intensity changing scenario. Finding the points of actual pressure applying and releasing can make a correlation with the blood hemoglobin concentration. This approach analyses a smartphone-based 30 s video photography of a nail bed, extracts feature and is finally able to predict the haemoglobin value for a human individual. The video photograph is captured within a fixed environmental condition so that for every individual no external interference has any impact on the video data. For this purpose, a dedicated pressure application device for the nail bed has been devised. This device provides a constant lighting illumination with a uniform and equal amount of pressure application for every individual irrespective of the size or thickness of the finger. The device consists of a base plate where the nail bed has to be placed and a white LED light is embossed just beneath the nail bed. The device is colored black from both inside and outside to prevent light emission from other sources and only the LED light rays from the LED light source can penetrate the nail bed. The event of pressure application and release induces the decolorization (deoxygenation) and recoloration (reoxygenation) of the tissues in the nail bed and the varying intensity of penetrated LED light rays through the nail bed can be observed.

This color-changing scenario is recorded through the smartphone camera module which is placed perpendicular at a distance to the top of the nail bed. The video photographs are captured encompassing the following criteria:

- ***initial stage:*** no pressure is applied on the nail bed, and the nail contains original nail color (reddish)
- ***pressure application:*** after 7 s of video recording for the initial stage, pressure is applied, and the color of the nail bed changes from red to pale
- ***pressure holding:*** after reaching the maximum pressure application point, the pressure is held for another 14 s
- ***pressure release:*** suddenly the pressure is released to release the stagnant blood in the nail bed tissue and the video photography is continued until the color of nail bed reaches its normal color (pale to pinkish to reddish)

Video preprocessing, ROI detection, pressure application and release point identification, feature extraction, and prediction model are described in the subsequent subsections.

3.1 ROI Selection

Initially, the captured video photographed is converted to frames having 30 frames per second. To detect the contour of the nail irrespective of the background, morphological contour detection approach is applied. This approach involves erosion, dilation, and a threshold gray level intensity value to distinguish the boundary line of the nail object. The colored image is converted to a binary image based on the threshold value, followed by the erosion property that disintegrates the boundaries of the foreground object and removes noises. Next, the dilation process increases the object area and connects the distorted boundary line. Finally, Freeman chain code is used to detect the four contour points that are sufficiently able to demarcate the shape of the object. This process traverses the boundary line and for each pixel, transcribes the direction it has traveled to reach the object. Based on these contour points, a rectangular region of interest of size 300×280 pixels are cropped from the entire nail-bed region. Figures 1(a), (b) and (c) show the original frame, contour detected frame and ROI selected frame, respectively.

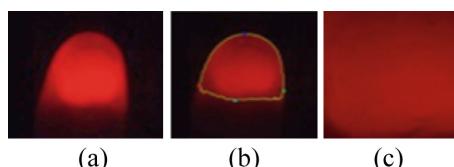


Fig. 1. (a) Original Frame extracted from video of Nail Pallor (b) Contour Detected Frame (c) ROI Detected Frame

3.2 Identification of Pressure Application and Release Point

The proposed method is established based on changing the color intensity value of the nail bed due to pressure application and release. Because of the deoxygenation and reoxygenation, the changes in the intensity level of successive frames are varying from time to time. It has been observed that the varying nature of the color intensity value is directly correlated to the blood haemoglobin concentration. To measure the rate of change of color, the Pearson Correlation coefficient is applied to compare each frame with the first frame which is considered as the reference frame. It measures the histogram spectrum differences between all the frames and the reference frame. Having similar histogram orientation of two frames provides a correlation value of 1, whereas significant changes decrease the correlation value up to -1 . Next, to identify the starting and stopping points of the pressure application (deoxygenation) and release (reoxygenation), gradient function is applied over the Pearson correlation coefficient values. It calculates the difference between each consecutive frame and thereby measures the maximum changes in the direction of the curve. The gradient function provides two maxima points that indicate the deoxygenation and reoxygenation stopping points respectively, whereas the two minima points denote the deoxygenation and reoxygenation starting points respectively. The characteristics of the color change due to the accelerated blood flow in nail bed tissue is shown in Fig. 2(b) that is experimentally derived from the Pearson correlation coefficient (PCC) values. The four red colored circles in Fig. 2(a) and Fig. 2(b) indicate the four critical points respectively.

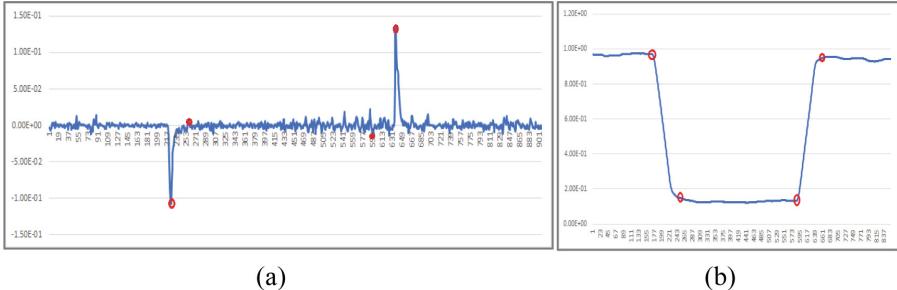


Fig. 2. (a) Gradient function over PCC and (b) Pearson Correlation Coefficient (PCC) curves

3.3 Feature Generation

As the proposed methodology is purely based on the color intensity values, three different color channels viz. Red Green Blue (RGB), Luma Chroma (YUV), and Hue Saturation Value (HSV) are considered for feature generation. For this approach, the importance of the four critical points is given as they reflect the maximum color changes in pallor. To propagate the features, only those frames where subtle changes in color intensity values are visible, are considered. The first ten frames, ten forward and ten backward frames each from four detected critical points are taken into account as they are the

most crucial frames and holding the information about the sudden changes in the color. For each of those selected frames, the mean intensity values of the R channel from the RGB channel, mean intensity values of the Y channel from the YUV channel, and mean intensity values of the V channel from the HSV channel are computed. Along with correlation values of those selected frames, age, and gender of the subjects are also taken into consideration. The present investigation observes that the varying nature of color changes are clearly visible only in the R channel, Y channel and V channel whereas the blue and green channel from RGB, UV from YUV, and HS from HSV do not contain much information, and hence, they are discarded. YUV colorspace is the affine transformation of RGB colorspace where Y correlates with the perceived intensity. HSV is helpful for lossy video compression, here the V channel relates with the brightness. Graphical representations of R, Y, and V channels are presented in Fig. 3. Finally, to predict the approximated haemoglobin concentration, a set of 362 features are generated from the video photography for a single subject.

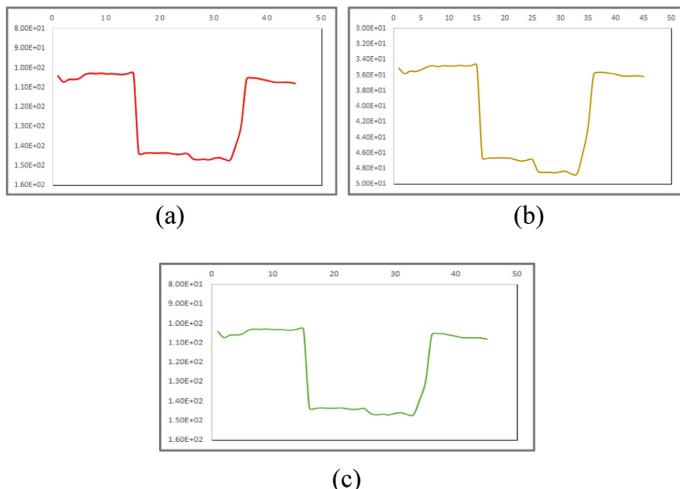


Fig. 3. Graphical representations of mean intensity values of each frame of (a) **R** Channel (b) **Y** Channel (c) **V** Channel

3.4 Dimensionality Reduction Using Autoencoder

Autoencoder is a multi-layer perceptron (MLP) neural network which follows the rule of backpropagation neural network. It comprises of two parts: encoder and decoder. A simple autoencoder consists of three layers: input layer, hidden layer, and output layer. There can be more than one hidden layer as per the user's need. The objective of an autoencoder is to regenerate the input vector at its outputs. That is an autoencoder tries to learn identity mapping. The network is trained to minimize the reconstruction error. In a single hidden layer autoencoder, input nodes take an input feature vector which is then mapped to the hidden space representation, called bottleneck, using a non-linear function.

The proposed work uses a single hidden layer autoencoder with 100 nodes, named as *GnAE*, having the capability to discretize the data points with an efficient way by adding Gaussian-neighborhood function with MSE function for optimization of reconstruction loss function. This Gaussian-neighborhood function $G(\cdot)$ helps to get similar data points closer in the latent space representation. To obtain a latent space feature representation, this autoencoder considers 362 features in the input layer and at the output layer it has been reconstructed. The latent space represents the reduced dimension of data, removing the irrelevant and multi-collinear data. This latent space representation is useful for learning the data features and for finding simpler representations of data for analysis. It represents the data in a compressed form where similar data points are close together. Figure 4 illustrates the architecture of Gaussian neighborhood based autoencoder.

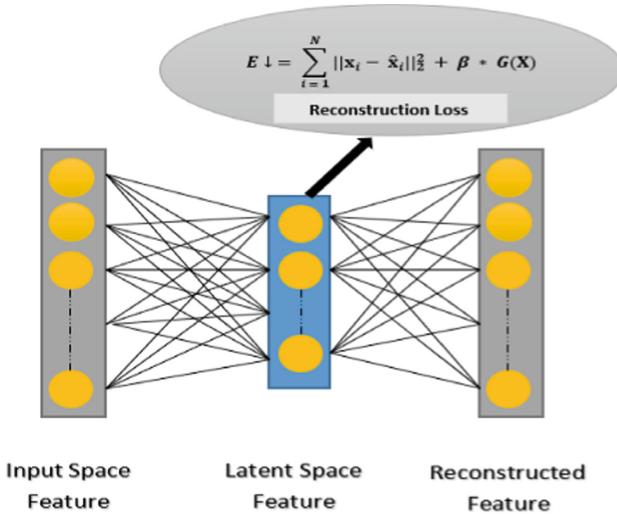


Fig. 4. Architecture of Gaussian-neighborhood based Autoencoder

The reconstruction loss, E of *GnAE* architecture is calculated in Eq. 1

$$E \downarrow = \sum_{i=1}^N \|x_i - \hat{x}_i\|_2^2 + \beta * G(X) \quad (1)$$

where \mathbf{X} denotes the dataset having N no. of samples and \mathbf{x}_i and $\hat{\mathbf{x}}_i$ represent original input and reconstructed output of i^{th} data respectively. Further, β is the regularizer and $G(\cdot)$ is given by the Eq. 2

$$G(X) = \sum_{i=1}^N e^{\frac{eqc^2}{2\sigma(t)^2}} \quad (2)$$

where eqc^2 denotes the Euclidean distance between the i^{th} data point to other points in dataset and $\sigma(t)$ is the radius of the neighborhood function which ensures the contribution of the neighborhood points on the i^{th} data point. The spread of $\sigma(t)$ gradually shrinks

as the iteration is increased. The $\sigma(t)$ is an exponentially decaying function defined in Eq. 3 as

$$\sigma(t) = \sigma_0 * e^{-\frac{t}{T}} \quad (3)$$

where σ_0 represents the initial radius, t is the current iteration and T is the total number of iterations.

3.5 Prediction Model

In the proposed scheme, the haemoglobin levels are predicted with the help of an autoencoder along with a combination of several regression models. The latent space representations, generated by autoencoder, now become more convenient to estimate the haemoglobin level. In our proposed model, eight regression models viz. Support Vector Regression (SVR) [15], Decision Tree Regression (DTR) [16], Ridge Regression (RR) [17], Bayesian Ridge Regression (BRR) [18], K-Nearest Neighbors Regression (KNNR) [19], Random Forest Regression (RFR) [20], Light Gradient boosting Machine Regression (LGBMR) [21], and Gradient Boosting Regression (GBR) [22] are considered. Regression models are the predictive modeling approach that analyses the interrelationship between the dependent and independent variables and draws the best fit line from where the distance between each data point is minimized. Here, the clinically validated haemoglobin levels are considered as dependent variables whereas the latent space feature representations are taken as independent or input variables. These non-linear regression models find the best fit curves or hyper-planes for minimizing the errors within an acceptable range. The latent space feature representations, generated from Autoencoder, are employed as the input to each of the above stated eight regression models. Out of those eight regression models, three best regression models are selected dynamically based on the RMSE values, to produce the final output. A weighted sum rule is applied over the predicted output generated from the best three regression models to predict the final output i.e., haemoglobin level. Increasing or decreasing the size of the dataset has no impact on the final prediction model, as the best three models are chosen dynamically based on the diversity of the data. Further, the outcome of the weighted sum rule neither contains much redundant information nor too little information, hence, the weighted sum rule becomes more advantageous in score level prediction. The weighted sum rule can be devised as.

$$\text{weighted_sum} = wt_1 \times P_1 + wt_2 \times P_2 + wt_3 \times P_3 \quad (4)$$

where $wt_1 = P_1/(P_1 + P_2 + P_3)$, $wt_2 = P_2/(P_1 + P_2 + P_3)$, $wt_3 = P_3/(P_1 + P_2 + P_3)$ and P_1, P_2, P_3 represent the predicted output of the best three regression models.

4 Experimental Result

To conduct the proposed experiment, real-time data from different health care organizations and hospitals are collected. This experiment is performed by 50 samples among which 11 are male samples and 39 are female samples with ages ranging from 19 to

62 years. For each of the subjects, haemoglobin concentration in blood is clinically evaluated and considered as the gold-standard. This entire procedure is conducted by obtaining the prior approval from the Institutional Ethical Committee as well as the written consent of the subject participating in the study. The range of the clinically tested haemoglobin value is from 7.7 to 14.1 g/dL. Data are collected from every individual using our dedicated nail device, maintaining the constant lighting illumination condition and camera focus.

To compare the proposed approach with other variants of autoencoder, different architectures, such as single layer autoencoder (AE), deep autoencoder (Deep AE) along with Gaussian-neighborhood function added in single layer autoencoder (GnAE) are considered. Using the cross-validation approach we have selected the optimal values for no. of the hidden layers, no. of nodes in the hidden layer, learning rate, value of σ and β of the GnAE. Table 1 represents a comparison result of the mean RMSE errors based on the leave-one-out cross-validation approach for different regression models where the input features are generated from different variants of the autoencoder. The term **Wt_Result** in the Table 1 represents the result of the weighted sum rule applied over the predicted output of the best three regression models.

Table 1. Mean RMSE of all prediction approaches

Regression model	Input space	Latent space		
	Original features	Features from AE	Features from DeepAE	Features from GnAE
SVR [15]	0.94	0.92	0.78	0.93
KNN [19]	0.81	0.72	0.83	0.82
Random Forest [20]	0.78	0.66	0.82	0.60
GBR [22]	0.63	0.65	0.79	0.59
Bayesian Ridge [18]	2.34	0.72	0.93	0.74
Ridge [17]	2.32	0.73	1.05	0.74
LGBM [21]	1.08	1.01	0.91	1.17
Decision Tree [16]	0.84	0.84	0.98	0.62
Wt_Result	0.70	0.61	0.76	0.56

The cross-validated architecture for single layer autoencoder is 362-100-362 where the latent space contains 100 feature values. The input and output layers have 362 nodes. This architecture is trained with learning rate of 0.05 for 1000 iterations. By increasing the hidden layers, the deep architecture is formed. The size of this architecture is 362-250-150-100-150-250-362 and trained with learning rate of 0.001 with 500 iterations. For the proposed GnAE architecture, simple single layer architecture same as AE is

followed where reduced dimension is 100, constructed from the 362 original features. This model is 500 times iterated with a learning rate of 0.05 for training purpose. For this model, the initial value of β is taken as 0.001 and the initial radius is considered as 20 which further decreased as the iteration is increased. In Table 1, it is observed that the proposed GnAE based features are more convenient than the other variants of autoencoder as well as the input space features. Though the deep architecture learns more complex data structure and track more information, however in the proposed model an increase in the number of hidden layers, decreases the efficacy of the model. As dataset size is small, increasing the number of hidden layer leads to overfitting the model and losing some needful information by unnecessarily squeezing the feature size and reduced the accuracy.

It is observed from the Table 1, that the weighted sum rule-based regression model gives better result as compared other state-of-the-art regression models. It is also remarkable that the weighted sum-based result for GnAE features provides the best result among all others.

The mean RMSE error is also compared with other state-of-the-art-methods stated in [6, 13] and [14] for haemoglobin estimation. They achieve RMSE value of 1.144, 1.35 and 0.6633 respectively. However, the proposed approach ensures an RMSE value of 0.56 which is significantly better than the existing solutions. Figure 5 illustrates the comparison of mean RMSE of the proposed method and the state-of-the-art methods.

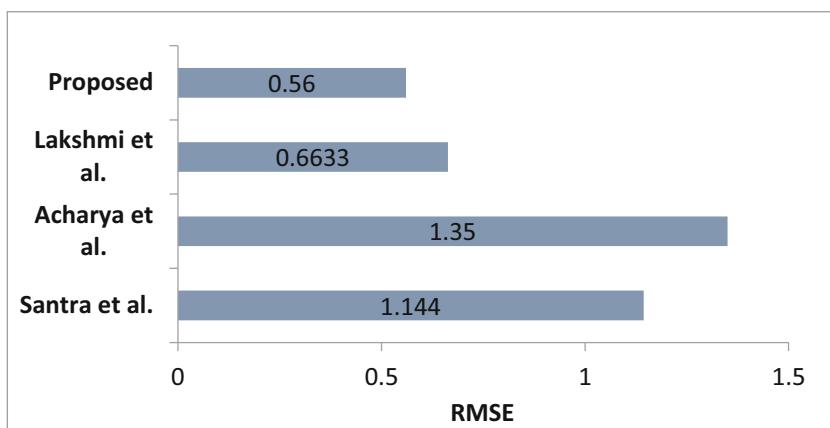


Fig. 5. Comparison of Mean RMSE error of the proposed and state-of-the-art methods

The predicted haemoglobin level of using weighted sum-based rule for the proposed model as well as other models and it's corresponding clinically evaluated haemoglobin values for each subject are plotted in the Fig. 6.

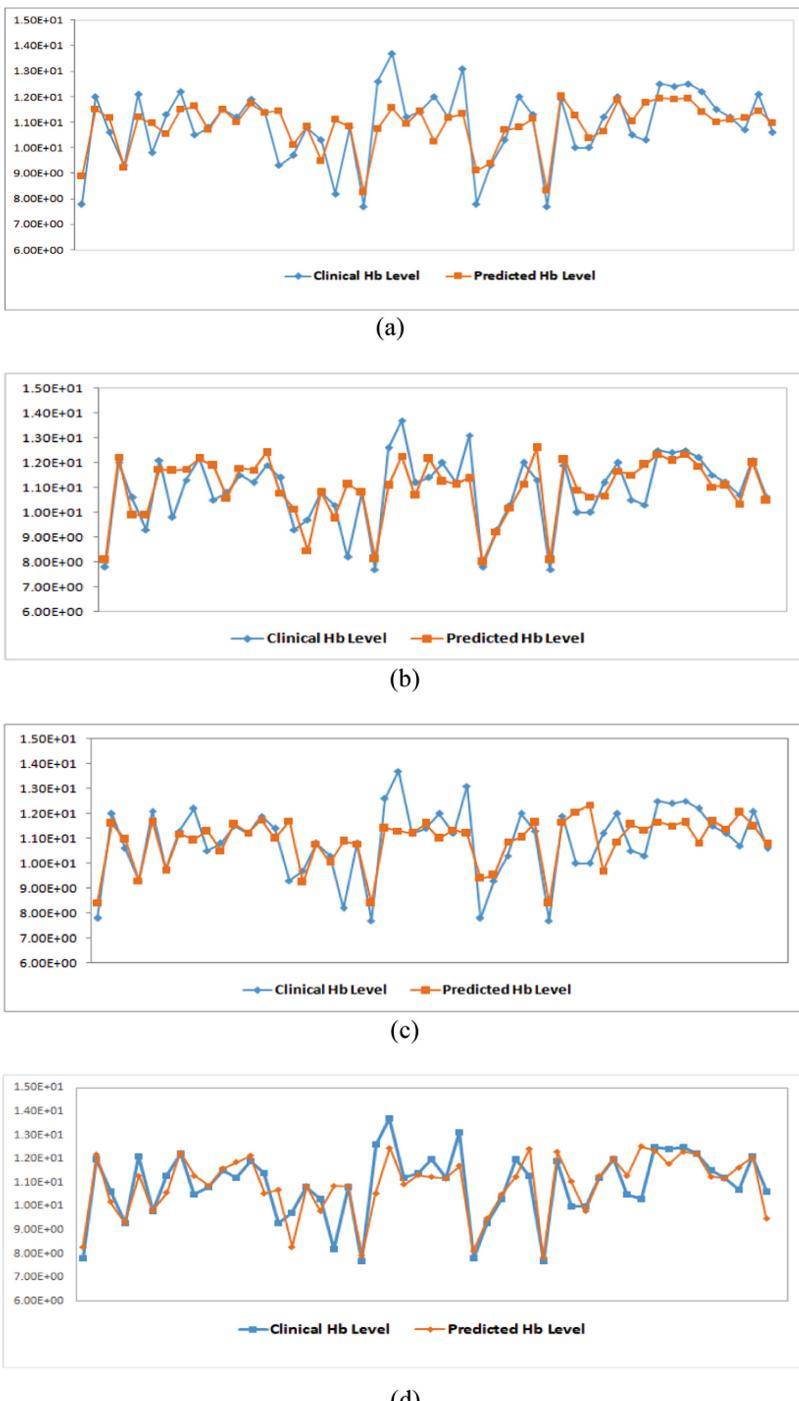


Fig. 6. Predicted Haemoglobin Level vs Clinical Haemoglobin Level extracted from weighted sum based regression model where features are (a) Input space (b) Single layer Autoencoder (c) Deep Autoencoder (d) Gaussian-neighborhood based Autoencoder

5 Conclusion

This paper has presented a non-invasive, instant haemoglobin level prediction system, with a minimal costly device dedicated for nail, by analyzing the color of the nail of the hand. The proposed method mainly analyses the color spaces of the videos of nail as the color of the nail bed changes from time-to-time due to occlusion and release of blood flow to the nail-bed. It achieves a mean RMSE of 0.56 which is lesser as compared to other state-of-the-art methods. The experiment is conducted in a real time environment. Further, there is a scope to improve the accuracy of the model by ensuring variety of haemoglobin samples.

Acknowledgment. We sincerely thank Dr. Dipankar Chakrabarti, Senior Physician, for providing us with preliminary level of clinical perspectives required for the work. We express our gratitude to PTMO, Parulia Health Centre, Durgapur and Superintendent, ESI Hospital, Durgapur, for their cooperation in the data collection process during the clinical study. The work is supported under a project sponsored by MeitY, GoI (Sanction number: 4(3)/2018-ITEA).

References

1. Muhe, L., Oljira, B., Degefu, H., Jaffar, S., Weber, M.W.: Evaluation of clinical pallor in the identification and treatment of children with moderate and severe anaemia. *Trop. Med. Int. Health* **5**(11), 805–810 (2000)
2. Ranganathan, H., Gunasekaran, N.: Simple method for estimation of hemoglobin in human blood using color analysis. *IEEE Trans. Inf. Technol. Biomed.* **10**(4), 657–662 (2006)
3. Kavsaoglu, A.R., Polat, K., Hariharan, M.: Non-invasive prediction of hemoglobin level using machine learning techniques with the PPG signal's characteristics features. *Appl. Soft Comput.* **37**, 983–991 (2015)
4. Atique, M.U., Sarker, Md. R.I., e Rabbani, K.S.: Measurement of haemoglobin through processing of images of inner eyelid. *Bangladesh J. Med. Phys.* **8**, 7–15 (2015)
5. Roychowdhury, S., Sun, D., Bihis, M., Ren, J., Hage, P., Rahman, H.H.: Computer aided detection of anemia-like Pallor. In: EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2017)
6. Santra, B., Mukherjee, D.P., Chakrabarti, D.: A non-invasive approach for estimation of hemoglobin analyzing blood flow in palm. In: 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, Melbourne (2017)
7. Tamir, A., et al.: Detection of anemia from image of the anterior conjunctiva of the eye by image processing and thresholding. In: Region 10 Humanitarian Technology Conference (R10-HTC). IEEE (2017)
8. Collings, S., Thompson, O., Hirst, E., Goossens, L., George, A., Weinkove, R.: Non-invasive detection of anaemia using digital photographs of the conjunctiva. *PLoS One* (2016). <https://doi.org/10.1371/journal.pone.0153286>
9. Mannino, R.G., et al.: Smartphone app for noninvasive detection of anemia using only patient-sourced photos. *Nat. Commun.* <https://doi.org/10.1038/s41467-018-07262-2>
10. Wang, E.J., Li, W., Hawkins, D., Gernsheimer, T., Norby-Slycord, C., Patel, S.N.: HemaApp: non invasive blood screening of hemoglobin using smartphone cameras. In: ACM UBICOMP 2016, Heidelberg, Germany (2016)

11. Magdalena, R., Saidah, S., Ubaidah, I.D.S., Fuadah, Y.N., Herman, N., Ibrahim, N.: Convolutional neural network for anemia detection based on conjunctiva palpebral images. *Jurnal Teknik Informatika (JUTIF)* **3**, 349–354 (2022)
12. Naik, B.R., Mude, S., Vennela, D.: Non-invasive Measurement of Hemoglobin for Rural India using Artificial Intelligence Algorithms (Preprint)
13. Acharya, et al.: Non-invasive estimation of hemoglobin using a multi-model stacking regressor. *IEEE J. Biomed. Health Inf.* **24**(6), 1717–1726 (2020)
14. Lakshmi, M., Manimegalai, P.: Non-invasive estimation of haemoglobin level using PCA and artificial neural networks. *Open Biomed. Eng. J.* **13**, 114–119 (2019)
15. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995). <https://doi.org/10.1007/BF00994018>
16. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986). <https://doi.org/10.1007/BF00116251.S2CID189902138>
17. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
18. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. Wiley, Hoboken (1973). ISBN 0-471-57428-7
19. Hodges, F.E., Joseph L.: Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas (1951)
20. Ho, T.K.: Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp. 278–282 (1995)
21. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., Stiglic, G.: Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **10**(1), 1–12 (2020)
22. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine (1999)

Author Index

- Alanezi, Ahmad III-277
Alatrany, Abbas Saad III-129
Alatrany, Saad S. J. III-129
Alejo, R. I-169, III-67
Al-Jumaili, Zaid III-277
Al-Jumaily, Dhiya III-129
Al-Jumeily, Dhiya III-220, III-277
Anandhan, Vinutha II-289
- Bai, Yunyi III-698, III-709
Bao, Wenzheng II-680, II-687, II-715, II-731
Basseur, Matthieu I-125
Bassiouny, Tarek III-277
Berloco, Francesco III-242
Bevilacqua, Vitoantonio III-242
Bi, Xiaodan I-701, II-103, II-260
Bin, Yannan II-757
Blacklidge, Rhys III-394
- Cai, Junchuang I-27
Cao, Dehua III-463
Cao, Yi II-663, II-670, II-697, II-705
Castorena, C. M. I-169
Cervantes, Jair I-391
Cervantes, Jared I-391
Chai, Jie III-341
Chang, Feng II-374
Chen, Baitong II-663, II-670, II-680, II-687, II-697, II-705, II-722
Chen, Cheng II-153
Chen, Debao I-112
Chen, Guang Yi I-330, I-420
Chen, Guanyuan II-356
Chen, Guolong I-292
Chen, Jianyong I-27, I-41
Chen, Jiazi II-663, II-670, II-697, II-705
Chen, Junxiong I-673
Chen, Mingyi I-535
Chen, Peng I-753, I-772, I-787
Chen, Shu-Wen II-588
Chen, Wen-Sheng I-267
Chen, Xiangtao III-288
Chen, Ying III-198, III-729
Chen, Yuanyuan III-802
- Chen, Yuehui II-334, II-394, II-615, II-663, II-697, II-705
Chen, Zhang I-245
Chen, Zhan-Heng I-739, II-220, II-451
Chen, Zhenqiong II-374, II-383
Cheng, Honglin II-680, II-687, II-715, II-739
Cheng, Li-Wei I-726
Cheng, Long III-755
Cheng, Meili I-701
Cheng, Zhiyu I-444, III-150
Choraś, Michał III-257
Chou, Hsin-Hung I-726
Chu, Jian II-731
Chu, Po-Lun I-726
Cloutier, Rayan S. II-588
Colucci, Simona III-242
Cong, Hanhan II-663, II-670, II-697, II-705
Cong, Hongshou I-772
Cuc, Nguyen Thi Hoa III-544
Cui, Xinchun I-412
Cui, Xiuming I-412
- Dai, Lai I-673
Dai, Yuxing I-430, I-494, II-569, II-579
Dai, Zhenmin I-379
Dai, Zichun II-319
Dalui, Mamata I-811
Das, Sunanda I-811
del Razo-López, F. III-67
Ding, Bowen I-68
Ding, Pingjian II-153
Ding, Wenquan II-517
Dmitry, Yukhimets III-504
Dong, Chao II-757
Dong, Chenxi I-401, III-117
Dong, Yahui III-380
Du, Jianzong I-412
Du, Yanlian I-51
Duan, Hongyu II-345
- Fan, Jingxuan I-506
Fan, Wei III-3
Fang, Ailian III-141
Fang, Chujie II-196

- Fang, Liang-Kuan III-626
 Fang, Yu I-339
 Feng, Cong III-604, III-662
 Feng, Yue I-412
 Feng, Zejing I-51
 Feng, Zizhou II-722
 Fengqi, Qiu III-106
 Filaretov, Vladimir III-55, III-93
 Fu, Qiming III-234
 Fu, Wenzhen II-670
- Gan, Haitao II-53
 Gangadharan, Sanjay II-289
 Gao, Guangfu II-628
 Gao, Lichao III-729
 Gao, Peng III-626
 Gao, Pengbo II-319
 Gao, Wentao II-28
 Gao, Xiaohua I-701
 Gao, Yun III-353
 García-Lamont, Farid I-391
 Ge, Lina I-638, III-802
 Ghali, Fawaz III-183
 Ghanem, Fahd A. III-304
 Gong, Xiaoling III-32
 Gong, Zhiwen I-444, III-209
 Gopalakrishnan, Chandrasekhar II-116, II-289, III-383
 Granda-Gutierrez, E. E. I-169, III-67
 Gu, Yi III-719, III-755
 Guan, Pengwei II-747
 Guan, Shixuan II-310
 Guan, Ying III-18
 Gubankov, Anton III-93
 Guo, Haitong I-13, II-41
 Guo, Jiamei III-198
 Guo, Xiangyang III-267
 Gupta, Rachna I-739
 Gurrib, Ikhlaas III-589
- Ha, CheolKeun III-80
 Ha, Cheolkeun III-544
 Han, Mengmeng II-757
 Han, Pengyong II-103, II-116, II-181, II-207, II-260, II-289, II-374, II-383, II-405, II-556
 Han, Tiaojuan I-3, I-13, II-41
 Han, Xiulin I-306
 Hao, Bibo I-739
 Hao, Zongjie III-463
- Haoyu, Ren III-106
 Harper, Matthew III-183
 He, Chunlin III-3
 He, Hongjian II-777
 He, Zhi-Huang I-317
 Hesham, Abd El-Latif II-14
 Hong, Sun-yan I-160
 Hsieh, Chin-Chiang I-726
 Hsieh, Sun-Yuan I-726
 Hu, Fan III-741
 Hu, Jiaxin II-715
 Hu, Jing II-496, II-507, II-517, II-533
 Hu, Juxi II-739
 Hu, Lun I-739, II-220, II-451
 Hu, Peng-Wei II-451
 Hu, Pengwei I-739
 Hu, Rong III-473
 Hu, Xiangdong I-627
 Hu, Yunfan I-149
 Hu, Zhongtian III-234
 Huang, Dingkai II-777
 Huang, Huajuan I-80, I-97, III-769, III-785, III-860
 Huang, Jian II-415
 Huang, Jiehao I-549
 Huang, Kuo-Yuan I-726
 Huang, Lei I-306
 Huang, Qinhuai III-162
 Huang, Youjun III-729
 Huang, Ziru II-415
 Huang, Ziyi II-28
 Hussain, Abir III-129, III-220
 Hussain, Abir Jaafar III-277
- Il, Kim Chung III-55
- Ji, Cun-Mei II-166, II-245, III-639
 Ji, Xuying I-112
 Jia, Baoli II-394
 Jia, Jingbo II-638
 Jiang, Kaibao II-356, II-364
 Jiang, Peng III-463
 Jiang, Tengsheng II-302, II-310
 Jiang, Yizhang III-741
 Jiang, Yu II-79
 Jianzhou, Chen III-106
 Jiao, Erjie III-198
 Jiao, Qiqi II-79
 Jin, Bo I-401, III-117

- Jing, Qu I-617
 Joshi, Rajashree I-739
- Kamalov, Firuz III-589
 Kang, Hee-Jun III-518, III-529
 Kesarwani, Abhishek I-811
 Khan, Md Shahedul Islam III-170
 Khan, Wasiq III-183, III-220, III-277
 Kisku, Dakshina Ranjan I-811
 Koepke, David I-739
 Komuro, Takashi I-472, I-483
 Kozik, Rafal III-257
 Krzyzak, Adam I-330
 Kuang, Jinjun I-564, I-589
- Lai, Jinling II-767
 Lai, Zihan II-807
 Lam, Dinh Hai III-544
 Le, Duc-Vinh III-80
 Le, Tien Dung III-518, III-529
 Lee, Hong-Hee III-484
 Lee, Mark III-394
 Lei, Peng II-66
 Lei, Yi II-507
 Lei, Yuan-Yuan III-626
 Leng, Qiangkui III-198
 Li, AoXing II-547
 Li, Bin II-166, III-18
 Li, Bo II-460, II-470
 Li, Dong-Xu II-451
 Li, Dongyi II-233
 Li, Feng II-345
 Li, Gang III-44
 Li, Guilin II-569
 Li, Hui I-137, I-196, I-209, I-221, II-628
 Li, Ji III-315
 Li, Jianqiang I-27, I-41
 Li, Jing I-579
 Li, Jinxin II-356, II-364
 Li, Juan III-380
 Li, Jun III-341
 Li, Junyi II-79
 Li, Lei II-166
 Li, Liang I-306
 Li, Lin II-405
 Li, Pengpai II-3
 Li, Qiang III-409
 Li, Renjie III-353
 Li, Ruijiang I-306
 Li, Shi I-181
- Li, Wenqiang I-535
 Li, Xiaoguang II-138, II-207
 Li, Xiaohui II-556
 Li, Xin I-209
 Li, Xin-Lu III-626
 Li, Yan II-345
 Li, Yanran II-116, II-289
 Li, Ya-qin I-160
 Li, Yaru III-18
 Li, Yuanyuan II-196
 Li, Zhang III-106
 Li, Zhaojia I-685
 Li, Zhengwei II-181, II-207, II-289, II-405
 li, Zhengwei II-383
 Li, Zhipeng III-304
 Lian, Jie II-569, II-579
 Liang, Yujun III-315
 Liang, Zhiwei III-846
 Lin, Fangze III-492
 Lin, Ke III-3
 Lin, Ning II-415
 Lin, Pingyuan I-430, I-494, II-569, II-579
 Lin, Qiuzhen I-27, I-41
 Lin, Xiaoli II-423, II-438, II-496, II-517, II-547
 Lin, Zeyuan I-673
 Ling, Ying III-891
 Liu, Bindong II-793
 Liu, Guodong III-409
 Liu, Hailei II-374, II-383
 Liu, Hao III-719
 Liu, Hongbo II-278
 Liu, Jie II-757
 Liu, Jin-Xing II-345
 Liu, Juan III-463
 Liu, Jun III-150, III-209
 Liu, Junkai II-302, II-310
 Liu, Qilin III-615
 Liu, Ruoyu III-719
 Liu, Shuhui II-126
 Liu, Si III-604, III-662
 Liu, Tao I-506
 Liu, Xiaoli I-412
 Liu, Xikui II-345
 Liu, Xiyu II-233
 Liu, Xujie I-430, I-494
 Liu, Yonglin I-412
 Liu, Yujun II-715
 Liu, Yunxia III-604, III-672
 Liu, Yuqing III-198

- Liu, Zhi-Hao III-639
 Liu, Zhi-Ping II-3
 Liu, Zhiyang III-448
 Lu, Jianfeng I-3, I-13, I-673, I-685, II-28, II-41, III-814
 Lu, Kun I-772, I-787
 Lu, Lei I-245
 Lu, Wang III-106
 Lu, Weizhong III-234
 Lu, Xingmin III-423
 Lu, Xinguo II-356, II-364
 Lu, Xinwei II-777
 Lu, Yaoyao II-302, II-310
 Lu, Yonggang III-18
 Luna, Dalia I-391
 Luo, Hanyu II-153
 Luo, Huaichao II-415
 Luo, Lingyun II-153
 Luo, Qifang III-830, III-846, III-860, III-876, III-891
 Luo, YiHua I-444
 Luso, Jiawei II-807
 Lv, Gang I-522
 Lv, Jiaxing I-772, I-787
 Lyu, Yi II-556
- Ma, Fubo II-319
 Ma, Jinwen III-267
 Ma, Zhaobin I-68
 Ma, Zuchang I-522
 McNaughton, Fiona I-739
 Mei, Jing I-739
 Meng, Qingfang II-334, II-394
 Meng, Tong II-705
 Mi, Jian-Xun III-353
 Min, Xiao I-363
 Min, Xu I-739
 Ming, Zhong I-27, I-41
 Miranda-Piña, G. I-169, III-67
 Moussa, Sherif III-589
- Nagahara, Hajime I-472
 Nazir, Amril III-589
 Nguyen, Duy-Long III-484
 Ni, Jian-Cheng II-166, II-245
 Nian, Fudong I-522
 Nie, Ru II-181
 Ning, Wei III-492
 Niu, Mengting II-14
- Niu, Rui II-270
 Niu, Zihan III-331
 Ouyang, Weimin III-162
 Pan, Binbin I-267
 Pan, Yuhang I-258
 Pang, Baochuan III-463
 Paul, Meshach II-289
 Pawlicki, Marek III-257
 Pei, ZhaoBin III-684
 Peng, Yanfei I-181
 Ponnusamy, Chandra Sekar II-289
 Premaratne, Prashan III-394
 Protcenko, Alexander III-55
 Pu, Quanyi I-352
- Qi, Miao I-506, III-380
 Qi, Rong II-245
 Qian, Bin III-473
 Qian, Pengjiang III-698, III-709, III-741
 Qiao, Li-Juan III-639
 Qiu, Shengjie I-535, I-549
 Qiu, Zekang I-663
 Qu, Qianhui III-719
- Ramalingam, Rajasekaran II-116, II-289, II-383
 Rendón, E. I-169, III-67
 Renk, Rafał III-257
- Sang, Yu I-181
 Sarem, Mudar I-535, I-564, I-589
 Shan, Chudong I-663
 Shan, Wenyu II-153
 Shang, Junliang II-345
 Shang, Li I-456, I-464
 Shang, Xuequn II-126, II-270, III-170
 Shao, Wenhao II-722
 Shao, Zijun II-722
 Shen, Yijun I-51
 Shen, Zhen II-767
 Sheng, Qinghua I-412
 Shi, Chenghao III-830
 Shi, Yan III-860
 Song, Wei III-423
 Su, Xiao-Rui II-451
 Su, Yanan III-719
 Sui, Jianan II-697
 Sun, Fengxu II-356, II-364

- Sun, Feng-yang III-654
 Sun, Hongyu II-556
 Sun, Hui III-380
 Sun, Lei I-444, III-209
 Sun, Pengcheng I-549
 Sun, Qinghua II-650
 Sun, Shaoqing I-306
 Sun, Yining I-522
 Sun, Zhan-li I-456, I-464
 Sun, Zhensheng II-345
 Sun, Zhongyu I-579
 Szczepański, Mateusz III-257
- Taleb, Hamdan III-304
 Tan, Ming III-435
 Tan, Xianbao II-92
 Tang, Daoxu II-364
 Tang, Yuan-yan III-435
 Tang, Zeyi I-430, I-494
 Tang, Zhonghua III-830
 Tao, Dao III-785
 Tao, Jinglu II-423
 Tao, Zheng II-687
 Tian, Wei-Dong III-341, III-367
 Tian, Yang II-470
 Tian, Yu I-245
 Tian, Yun II-405
 Topham, Luke III-220
 Tran, Huy Q. III-544
 Tran, Quoc-Hoan III-484
 Truong, Thanh Nguyen III-518, III-529
 Tsunezaki, Seiji I-483
 Tun, Su Wai I-472
 Tun, Zar Zar I-483
- Valdovinos, R. M. I-169, III-67
 Van Nguyen, Tan III-544
 Vladimir, Filaretov III-504
 Vo, Anh Tuan III-518, III-529
 Vu, Huu-Cong III-484
- Wan, Jia-Ji II-588
 Wang, Bing I-753
 Wang, Chao I-292
 Wang, Chaoxue I-258
 Wang, Dian-Xiao II-166
 Wang, Dong II-747
 Wang, Guan II-650
 Wang, Haiyu I-653
 Wang, Han I-277
- Wang, Hongdong II-722
 Wang, Hui-mei I-160
 Wang, Jia-Ji II-600
 Wang, Jian II-116, II-615, III-32
 Wang, Jian-Tao III-435
 Wang, Jianzhong I-506
 Wang, Jing I-412
 Wang, Jin-Wu I-379
 Wang, Kai I-317
 Wang, Qiankun II-181
 Wang, Ruijuan I-221
 Wang, Shaoshao I-277
 Wang, Shengli I-430, I-494
 Wang, Weiwei II-278
 Wang, Wenyan I-772, I-787
 Wang, Xiao-Feng I-317, III-435
 Wang, Xue III-814
 Wang, Xuqing III-719
 Wang, Yadong II-79
 Wang, Ying III-684
 Wang, Yingxin I-673
 Wang, Yonghao III-672
 Wang, Yuli III-234
 Wang, Yu-Tian II-166, II-245, III-639
 Wang, Zhe I-638
 Wang, Zhenbang I-258
 Wang, Zhipeng II-207
 Wang, Zhuo II-722, II-731
 Waraich, Atif III-220
 Wei, Xiuxi I-80, I-97, III-769, III-785, III-876
 Wei, Yixiong III-331
 Wei, Yuanfei III-891
 Wei, Yun-Sheng I-317
 Weise, Thomas III-448
 Win, Shwe Yee I-483
 Wu, Daqing III-267
 Wu, Geng I-401, III-117
 Wu, Hongje III-615
 Wu, Hongjie I-352, I-799, II-66, II-92, II-302, II-310, III-234, III-304
 Wu, Lijun II-757
 Wu, Lin II-415
 Wu, Mengyun II-460
 Wu, Peng II-615, II-628, II-638
 Wu, Qianhao III-331
 Wu, Shuang III-367
 Wu, Xiaoqiang I-41
 Wu, Xu II-747
 Wu, Yulin II-650

- Wu, Zhenghao II-438
 Wu, Zhi-Ze I-317
 Wu, Zhize III-448
 Wu, Ziheng I-772, I-787
 Xia, Junfeng II-757
 Xia, Luyao I-673, I-685
 Xia, Minghao II-496
 Xiahou, Jianbing I-494
 Xiang, Huimin II-547
 Xiao, Di III-463
 Xiao, Kai II-680
 Xiao, Min III-170
 Xiao, Ming II-319
 Xie, Daiwei I-379
 Xie, Jiang II-777, II-793
 Xie, Kexin I-267
 Xie, Lei I-292
 Xie, Wen Fang I-420
 Xie, Wenfang I-330
 Xie, Yonghui I-221
 Xie, Zhihua III-409
 Xieshi, Mulin I-430, I-494
 Xin, Sun I-617
 Xin, Zhang III-106
 Xing, Yuchen I-137
 Xu, Caixia II-116, II-207, II-289, II-374, II-405
 Xu, Cong I-763
 Xu, Hui I-506, III-380
 Xu, Lang I-444
 Xu, Li-Xiang III-435
 Xu, Mang III-698, III-709
 Xu, Mengxia III-814
 Xu, Xian-hong III-654
 Xu, Xuexin I-430, I-494
 Xu, Youhong I-799
 Xu, Yuan I-535, I-549, I-564, I-589
 Xue, Guangdong III-32
 Xuwei, Cheng III-106
 Yan, Jun III-234
 Yan, Rui II-138
 Yan, Wenhui II-757
 Yang, Bin I-579, II-650
 Yang, Bo II-747
 Yang, Chang-bo I-160
 Yang, Changsong III-234
 Yang, Chengyun I-245
 Yang, Hongri II-334
 Yang, Jin II-556
 Yang, Jing III-435
 Yang, Jinpeng I-701, II-103, II-260
 Yang, Lei II-481
 Yang, Liu I-627
 Yang, Weiguo I-579
 Yang, Wuyi I-233
 Yang, Xi I-627
 Yang, Xiaokun II-207
 Yang, Yongpu II-53
 Yang, Yuanyuan III-473
 Yang, Zhen II-687
 Yang, Zhi II-53
 Yao, Jian III-741
 Ye, Lvyang III-769
 Ye, Siyi II-507
 Yin, Ruiying II-319
 Yixiao, Yu I-605
 You, Sheng II-807
 You, Zhu-Hong II-220, II-451
 You, Zhuhong II-270
 Yu, Changqing I-763
 Yu, Chuchu I-80
 Yu, Jun II-319
 Yu, Naizhao I-363
 Yu, Ning II-245
 Yu, Xiaoyong I-292
 Yu, Yangming I-401, III-117
 Yu, Yixuan III-876
 Yuan, Changgan I-352, I-799, II-66, II-92, III-55, III-93, III-304, III-504
 Yuan, Changgan III-615
 Yuan, Jianfeng II-747
 Yuan, Lin II-767
 Yuan, Qiyang I-339
 Yuan, Shuangshuang II-615
 Yuan, Zhiyong I-663
 Yukhimets, Dmitry III-93
 Yun, Yue II-270
 Yupei, Zhang II-126
 Zaitian, Zhang III-106
 Zeng, Anping III-3
 Zeng, Rong I-13, II-41
 Zeng, Rong-Qiang I-125
 Zha, Zhiyong I-401, III-117
 Zhai, Pengjun I-339
 Zhan, Zhenrun II-103, II-260
 Zhang, Aihua I-277

- Zhang, Bingjie III-32
 Zhang, Dacheng III-473
 Zhang, Fa II-138
 Zhang, Fan I-258
 Zhang, Guifen III-802
 Zhang, Hang II-533
 Zhang, Hao I-3, I-13, I-638, I-685, II-28, II-41, III-814
 Zhang, Hongbo I-306
 Zhang, Hongqi III-331
 Zhang, JinFeng I-444, III-209
 Zhang, Jinfeng III-150
 Zhang, Jun I-753, I-772, I-787
 Zhang, Kai II-638
 Zhang, Le II-319
 Zhang, Lei I-245
 Zhang, Le-Xuan III-626
 Zhang, Liqiang II-615
 Zhang, Lizhu III-44
 Zhang, Meng-Long II-220
 Zhang, Na II-747
 Zhang, Ping II-451
 Zhang, Qiang II-138, II-394
 Zhang, Qin I-627
 Zhang, Shanwen I-306, I-763
 Zhang, Tao I-233
 Zhang, Tian-Yu III-341
 Zhang, Tian-yu III-367
 Zhang, Tianze II-628
 Zhang, TianZhong III-209
 Zhang, Tingbao I-663
 Zhang, Wensong III-557, III-572
 Zhang, Wu II-793
 Zhang, Xiang I-430, I-494, II-356
 Zhang, Xiaolong II-423, II-438, II-481, II-496, II-533
 Zhang, Xiaozeng III-141
 Zhang, Xin I-68
 Zhang, Xing III-557, III-572
 Zhang, Xinghui I-753
 Zhang, Xinyuan II-517
 Zhang, Yan I-112
 Zhang, Yang II-79
 Zhang, Yanni I-506
 Zhang, Yin I-196
 Zhang, Yu I-233
 Zhang, Yuan I-739
 Zhang, Yue II-663
 Zhang, Yupei III-170
 Zhang, Yuze I-456, I-464
 Zhang, Yu-Zheng III-367
 Zhang, Zhengtao II-181
 Zhao, Bo-Wei I-739, II-220, II-451
 Zhao, Guoqing II-3
 Zhao, Hao I-522
 Zhao, Hongguo III-604, III-672
 Zhao, Jianhui I-663
 Zhao, Liang I-363
 Zhao, Meijie III-288
 Zhao, Tingting II-103, II-260
 Zhao, Wenyuan I-663
 Zhao, Xin II-747
 Zhao, Xingming I-352, I-799, II-66, II-92, III-304
 Zhao, Xinming III-615
 Zhao, Yuan I-772, I-787
 Zhao, Zhong-Qiu III-341, III-367
 Zhao, Ziyu II-507
 Zhaobin, Pei I-605, I-617
 Zheng, Aihua I-292
 Zheng, Chunhou I-753, II-138
 Zheng, Chun-Hou II-245, III-639
 Zheng, Dulei I-339
 Zheng, Jinping II-556
 Zheng, Xiangwei I-412
 Zheng, Yunping I-535, I-549, I-564, I-589
 Zheng, Yu-qing III-654
 Zheng, Zhaona II-650
 Zhirabok, Alexey III-55
 Zhong, Guichuang I-535
 Zhong, Ji II-638
 Zhong, Lianxin II-334
 Zhong, Yixin II-670
 Zhou, Hongqiao III-331
 Zhou, Jiren II-270
 Zhou, Yaya III-170
 Zhou, Yihan II-747
 Zhou, Yongquan III-802, III-830, III-846, III-860, III-876, III-891
 Zhou, Yue II-715
 Zhou, Zhangpeng I-430, I-494
 Zhu, Fazhan I-772, I-787
 Zhu, Hui-Sheng II-588
 Zhu, Jun-jun III-367
 Zhu, Qingling I-27, I-41
 Zhu, Xiaobo III-615
 Zhu, Yan I-160
 Zhu, Yumeng II-722
 Zhuang, Liying I-412
 Zitong, Yan III-106

- Zong, Sha III-409
Zou, Feng I-112
Zou, Le I-317
Zou, Pengcheng I-97
Zou, Quan II-14
Zou, Zhengrong III-492
Zou, Zirui I-549
Zu, Ruiling II-415
Zuev, Alexander III-55
Zuo, Zonglan II-138